

# DNS-SZEKVENCIÁK ANALÍZISE SZÖVEGELEMZÉSI MÓDSZEREKKEL

Készítette: **Bír Tams**  
V. ves fizikushallgat  
ELTE TTK

Tmavezető: **Vicsek Tams**  
egyetemi tanr  
ELTE TTK Biológiai Fizika Tanszk

**Diplomamunka**

**Budapest, 1998.**

## Összefoglaló

Az elmúlt néhány évtizedben mind a genetikában, mind a nyelvészetben hatalmas fejlődés következett be, amely lehetővé teszi azt, hogy a fizika, azon belül is a modern statisztikus fizika, új szempontokat és kutatási irányokat felvesszen a tudományok fejlődéséhez. A rohamosan növekvő számban hozzáférhető DNS-szekvenciák, ennek, sőt egyre több egész kromoszma és a teljes genom, nem csupán lehetővé teszi, de – mint azt a második fejezetben megmutatom – igényli is azt, hogy statisztikai eszközökkel elemezzék őket.

Dolgozatom harmadik fejezetében ismertetni fogok több modern megközelítést is (hosszú távú korrelációk, Zipf-analízis, információméleti entrópia és redundancia), amelyeket természetes nyelven írt szövegek és DNS-szekvenciák analízisére egyaránt alkalmaztak. E megközelítések arra az eredményre vezetnek, hogy a nem kódoló DNS-szakaszok sok tulajdonságukban hasonlítanak az írt szövegekre. Mivel kiderült, hogy Markov-folyamatokkal nem írható le kellő pontossággal ezek a statisztikus tulajdonságok, a sztochasztikus környezetfüggetlen grammatikák elméleti fogalmát javaslom a jelenségek leírására. Vgl. diplomamunkám negyedik fejezetében egy vektortriangulációs technikát alkalmaztam a DNS-szekvenciákra. A módszer alkalmasnak bizonyult írt szövegek nyelv és tartalom szerinti szétválogatásában, most a kódoló és a nem kódoló DNS-szakaszok elkülönítésére is sikerrel alkalmazzuk.

Szeretném ezen a ponton köszönetemet kifejezni tmavezetőmnek, *Vicsek Tamás* tanár-szervező egyetemi tanárnak, aki minden lépésben nagyban hozzájárult a dolgozat elkészítéséhez. Köszönet illeti *Major Gáspárt* is, aki az ELTE TTK Genetika tanszékéről a biológus szemszövegéből segített a munkámat, valamint *Czirk András* doktoranduszának, akitől a Damashek-módszer DNS-szekvenciákra történő alkalmazásának az elveit szívesen megtanultam, valamint sok további értékes tanácsot és szakirodalommal segített.

# Tartalomjegyzék

<b>1. Bevezets</b>	<b>4</b>
1.1. A statisztikus fizika interdiszciplnr is alkalmazsaibl	6
1.2. Biologia s statisztikus fizika	8
<b>2. A DNS-szekvencik kutatsnak irnyai</b>	<b>11</b>
2.1. A statisztikai megkzelts kt clja	11
2.2. A g nidentifikci mdszerei	14
<b>3. DNS-szekvencik s rott szvegek modellezse statisztikai eszkzkkel</b>	<b>17</b>
3.1. Markov-folyamatok	19
3.2. A Zipf-analzis	20
3.3. Hossz tv korrelcik	23
3.4. A klcsns informci-fggvny	28
3.5. A redundancia	30
3.6. Modellezs generativ nyelvtanok segtsgvel	31
3.6.1. A formlis nyelvek elmletnek alapfogalmai	32
3.6.2. Sztochasztikus nyelvtanok	37
3.6.3. Becsls a Zipf-fggvnyre	38
<b>4. A vektortr-technika</b>	<b>42</b>
4.1. A Damashek-mdszer ismertetse	43
4.2. A mdszer tovbbfejlesztse s diszkusszija rott szvegekre	46
4.3. A szmtsi algoritmus	49
4.4. A felhasznlt adatbzis	50
4.5. A vektortr-technika alkalmazsa DNS-szekvencikra	52
<b>5. sszefoglals</b>	<b>58</b>
<b>Irodalomjegyzk</b>	<b>61</b>

# 1. fejezet

## Bevezetés

”A clash of doctrines is not a disaster, it is an opportunity.” ■

Whitehead <sup>1</sup>

gy tűnik, a második világháború, ill. az azt megelőző, politikailag feszült korszak és az új gazdasági nehézségeket követően, az 1950-es években új lendület vehetett a tudomány. Jelenlegi diszciplínában forradalmi változások zajlottak le ezekben az években. A számítástechnika hőskorszaka és az úrkutatás kezdete (1957) egybeesik az első atomerőmű (1954) felépítésével, az antiproton (1955) és a paritsztrák (1956-57) felfedezésével. 1956-ban vezeti be Rindler az eseményhorizont fogalmát. És hogy teljesen más területekről is idézünk példákat, az ötvenes évek végén kezd kibontakozni a kognitív pszichológia, valamint az ötvenes évek közepén publikálják az 1947-től kezdve felfedezett Holt-tengeri tekercsek első fragmentumait, amelyek a Biblia-tudományban, az ókori Palesztina történetének a kutatásban és a héber nyelvészetben egyaránt váratlan fordulatot hoztak.

Bennünket most hrom, első pillanatra nagyon távolinak tűnő tudományok ezekben az években indult megjelenni. Időrendben haladva, elsőnek a *genetika* forradalma zajlott le: 1953-ban jelentette meg *J. D. Watson* és *F. H. C. Crick* azt a rövid cikket, amely nélkül nem képzelhetnénk el az elmúlt negyven év biológiáját. *A. N. Kolmogorov* 1954-es tételével (*V. I. Arnold* és *J. K. Moser* általánosításával, az *n. KAM-tétel*), a statisztikus fizikában nyitott új területeket, amely a kozmológiai elmélethez, a komplex rendszerek fizikájához, a fraktálok fizikai alkalmazásaihoz vezetett (Szepfalussy & Tl [1982], Mainzer [1997]). Végezetül, 1957-ben látott napvilágot *N. Chomsky* *Syntactic Structures* című műve, amely – függetlenül attól, hogy egyetértünk-e az előfeltevésével

---

<sup>1</sup> ”Ha a doktrínák csak egymással, az nem katasztrófa, hanem kihasználható esély.”, *Whitehead, Science and Modern World, p. 186, idzi Prigogine & Stengers [1986]*.

– a nyelvtudomány teljesen új megközelítést jelentette, a korábbi, strukturalista nyelvészettel szemben.

Mi hozza közs nevezőre ezt a három tudományt? Mindhárom tudományban olyan kérdéseket lehet a század kezdetétől kezdődően egzakt formában feltenni és a természettudományos modellalkotás-jelölést ellenőrző kutatási módszer segítségével megválaszolni, amelyek a megelőző századokban csupán a filozófiai spekuláció tárgyai lehettek.

A modern genetika véglegesen új területet nyitott több ezer a legjelentősebb elméleti elképzelésének: az ősi elképzelésnek, amely a vrhez kapcsolta az öröklődést,<sup>2</sup> valamint sok korai "tudományos", de nem megfigyelésre alapozott elképzelésnek (Berend M. [1980]), amelyeket rendre megcáfoltak. Ilyen volt például az *ősnemzés tana*, mely az isteni léleknek (ugyanis ezen elképzelés szerint a nemzés legfontosabb pillanata "a lélek beltetése a test anyagába") vagy a hímnek tulajdonította az utód tulajdonságait. *Paracelsus* (1493-1514) svájci természettudós és orvos tanai, amelyek szerint a terhes nő vágyai, lmaai, benyomásai jelennek meg a születendő gyermek tulajdonságaiban, míg hatnak a szélesebb társadalmi körtegek gondolkodásában. A XVII-XVIII. század *preformizmusa* hirdetői szerint pedig az ivarsejtekben *preformált homunkuluszok*, miniatűr emberkék találhatók, amelyeknek szintén vannak még kisebb homunkuluszokat tartalmazó ivarsejtjei, és így tovább. A genetika tudománya akkor jött létre, amikor a kísérletek már nem utalnak csupán a spekulatív elméletekre, hanem a kísérletek eredményeiből kiindulva alátámasztották fel őket. *G. J. Mendel* (1822-1884) törvényei, majd a XX. század első felének kutatásai (*H. de Vries*, *C. E. Correns*, *A. v. S. Tschermak*, *A. Weismann*, *O. T. Avery*, *T. H. Morgan* és sokan mások) tették lehetővé, hogy *Watson* és *Crick* felfedezése megalapozza az öröklődés igaz elméletét.

*Noam Chomsky* kvetői immáron szinte természettudománynak tekintik a nyelvészetet. A korábbi szemlélet szerint ez a tudomány eljárási a nyelvi rendszer leírása volt, és ezt – elméleti megalapozottságban, precizitásban – a strukturalisták vitték tovább a legteljesebb mértékben, a XX. század első felében. Chomsky első műve utal, hogy kvetjük iskoljának szintén változó, fejlődő gondolatait, hogy nem, a nyelvész feladata egy jóval mélyebb leírást adni és minél adekvátabb *modell* alkotása, amelyek eleget tesznek bizonyos, a modellel szemben támasztott, elméleti elvárásoknak is.<sup>3</sup> Az első modellek, amelyek csak a Newton-i mechanika természetmodellje, csupán egy nagyon szűk és idealizált jelenségre leírásra voltak alkalmasak, de a tudomány lassú fejlődése során szélesedett (és, reméljük, szélesedni fog) a leírható adathalmaz.

---

<sup>2</sup> Gondoljunk például a "körvonal", "vonal" vagy a "váltó" kifejezésekre.

<sup>3</sup> A modellnek egységes leírást kell adnia minél több nyelv minél változatosabb jelenségeiről, meg kell felelnie néhány nyelvészeti elvárásnak, figyelemmel kell lenni a gyermek nyelvvelajtságnál és affinis betegeknél megfigyelt jelenségekre, és pszichológiailag is adekvátnak kell lennie.

A chomskynus nyelvészet másik nagy eredménye a matematikai eszközök bevonása a modellekbe. Idézzük Chomsky szavait:<sup>4</sup> *”Figyeljék meg, hogy egybként az egyetemese nyelvten pontosan krlrt elveinek ltezse teszi lehetőv egy j terület, a matematikai nyelvészet létrejttt, mely az egyetemese nyelvtenban meghatározott megszoztorsoknak eleget tevő generatív rendszerek osztlyt veti elmleti vizsgálata alá. Ez a vizsgálat valamennyi lehetséges emberi nyelv formális sajtságnak részletes feltrásra trekszik. A terület mg gyerekcipőben jr; csak az utbbi vtizedben vlt egy ilyen vállalkozás lehetősége elkpzeltetőv. ... Úgy tűnik tehát, hogy a matematikai nyelvészet jelen pillanatban egyedllan kedvező helyzetben van a trsadalomtudományok s a pszichológia matematikai megkzeltsei kztt ahhoz, hogy ne egyszerűen adatok elmletv,<sup>5</sup> hanem az emberi mentális folyamatok jellegt meghatároz rendkvl elvont elvek s struktrk vizsgálatv fejlődjn.”*

Ezek a matematikai nyelvészeti modellek nem csupán a kognitív tudományokra vannak jelentős hatással, nem csupán egzakt formába ntették a korábbi fogalmakat, s nem is csak a nyelvtudomány tudományossági kritériumait vltoztatták meg, hanem potenciálisan lehetőséget teremtenek arra, hogy a nyelv kvantitatív viszonyait is beptsk a nyelvészeti modellekbe. Ezen a ponton kapcsolódik a nyelvészethez a fizika, mint eredetileg az (lettelen?) természet kvantitatív jelenségeit ler tudomány, s dolgozatom későbbi részeiben be fogom mutatni eme vizsgálatok első, főleg fizikusok ltal produkált eredményeit is.

## 1.1 A statisztikus fizika interdiszciplnris alkalmazásaiból

A genetikában s a nyelvészetben lezajlott vltozások után, most trjnk vissza az 1950-es évek nagy, tudománytrtneti fordulataihoz, azon belül is a fizika j gáihoz. A modern statisztikus fizika tette lehetőv az nszerveződő komplex struktrk vizsgálatát. A klasszikus mechanika, az időnyíl megfordíthatóság miatt, csupán a legegyszerűbb folyamatokat, pldül a bolygmozgást vagy az ingamozgást volt képes lerni. A múlt századi termodinamika irreverzibilis világa les ellentmondásba került a reverzibilis mechanikával, de továbbra is csak a homogén egyensúly, a ”hőhall” irányba trtnő vltozásokat tudta értelmezni. Csupán az elmélet vtizedek fejlődése eredményeként fogalmazódott meg az, hogy *”az irreverzibilitásnak a természetben alkot szerepe van, hiszen spontán szerveződési folyamatokat tesz lehetőv. A fizikán belül az irreverzibilis folyamatok tudománya visszallotta jogaiba a tevékenység- s nszaport struktrkat létrehoz természet*

---

<sup>4</sup> Chomsky [1968], magy. kiadás 227. o.

<sup>5</sup> Chomsky itt a kvantitatív nyelvészetre utal (ld. pl. Nagy F. [1986]-t), amely egyszerű statisztikai megfigyelésekből von le nem igazán messze mutat, helyenként erősen vitatható következtetéseket. (Itt pldül a glottokronológia kritikjra gondolok.)

*gondolatt.*” (Prigogine-Stengers [1986] p. 10.) Megnylt ezzel az út az összetett rendszerek fizikai eszközzel történő vizsgálatához, ahhoz, hogy az alkotó elemektől függetlenül, a struktúrát is lehessen elemezni.

Ugyanakkor a természetben előforduló összetett rendszerek vizsgálata, éppen a fizika korábbi gainak említett jellegéből adódóan, kiegészített a tudomány hagyományos területeiből. Ezért okoz sokaknak meglepetést, ha például biológiai, közgazdasági vagy szociológiai rendszerek vizsgálatba fizikusok is bekapcsolódnak. Pedig eme interdiszciplináris kutatások manapság egyre több fizikust vonzanak, és nagyon gyorsan tűnnek.

A nem-lineáris statisztikus fizika fedezte fel az 1980-as években azt, hogy egy sor természeti jelenség fraktálszerű képződményeket hoz létre a természetben (Family & Vicsek [1991], Vicsek [1992]). Ezt követően, a fraktálok és a kaotikus jelenségek kutatásának "divatja" terjedt a közadalmi tudományokba, sőt a művészetelméletbe is. A *Rend és Kosz* [1997] című, a közelmúltban megjelent tanulmánykötetben a közelmélet történelmi és közgazdasági alkalmazásai mellett, képzőművészeti és zenei alkotások fraktáltulajdonságokkal történő elemzésre is találunk kísérleteket.

A közgazdaságtanban egy sor jelenség elemzésnél merülhetnek fel a statisztikus fizikából kölcsönzött fogalmak. (További példákat hoz Mainzer [1997] pp. 270-281. is.) Itt nem is elsősorban a termodinamikai fogalmak és törvényszerűségek gazdasági következményeinek a vizsgálatra gondolok, amelyre például szolgálhat *Csekő . & K. Martins* (in: K. Martins, M. Moreau [1995]), valamint Martins [1997]. Az 1990-es évek közepén (jelen) a statisztikus fizikával foglalkozó kutatók rádeklődési körbe kerültek egyes skálási tulajdonságot mutató közgazdasági jelenségek és kaotikusnak tűnő folyamatok.

Az egyesített államokbeli ipari cégek másfél évtizedet tartó adatai alapján azt tapasztaltuk, hogy az azonos éves forgalommal rendelkező cégek között az éves logaritmikus növekedési rátá exponenciális eloszlást mutat (M. H. R. Stanley et al. [1996a,b]):

$$p(r|s_0) = \frac{1}{\sqrt{2}\sigma(s_0)} \exp\left(-\frac{\sqrt{2}|r - \bar{r}(s_0)|}{\sigma(s_0)}\right), \quad (1.1)$$

ahol  $s_0$  az éves  $S_0$  forgalom logaritmusának egy adott értékében,  $r$  pedig ennek változása a következő értékében ( $r = \ln(S_1/S_0)$ ). Az eloszlás  $\sigma(s_0)$  szórása csökken a növekvő forgalom mellett, mghozz meglepő skálási tulajdonsággal rendelkezik:

$$\sigma(s_0) = a \exp(-\beta s_0) = a S_0^{-\beta}. \quad (1.2)$$

A skálási exponens  $\beta = 0,15 \pm 0,03$ , ha az eladások változást vizsgáljuk, ill.  $\beta = 0,16 \pm 0,03$ , ha az alkalmazottak számát tekintjük a fentiekkel analóg módon. A jelenség magyarázatra izgalmas modellel szolgál Amaral et al. [1997].

Gazdasági mutatók (az ún. Standard & Poor's 500 index) valószínűségi eloszlásnak szelvési tulajdonságát mutatta meg Mantegna & Stanley [1995a], három nagyszámú kereskedelmi, az 1 perctől az 1000 percig tartó időskálán (ld. még Y. Liu et al. [1997]).

A rövid távú, spekulatív valutapiac jelenségeit Ghashghaie et al. [1996, 1997] (ld. még Mantegna & Stanley [1996, 1997]) a hidrodinamikai turbulenciához hasonlítja, mivel az előbbiben olyan "információ-kaszádokat" telez fel, amelyek a hidrodinamikai turbulencia energia-kaszádjainak felelnek meg.

Jean-Philippe Bouchaud és társainak munkássága az elmúlt években a pénzügyi kockázatok és az opciók részének területén hozott új eredményeket (Bouchaud & Potters [1997], Potters et al. [1997]): az opciók részének használatos, 1973-ban származó és 1997-ben közgazdasági Nobel-díjjal jutalmazott *Black – Scholes-formula* hinyosságaira mutatnak rá, mind elméleti számításokkal, mind valódi adatok feldolgozásával.

A statisztikus fizika közgazdasgtani és pénzügyi alkalmazásaiban hozott felsoroltak végén hadd említsünk egy magyar példát is: Várvai Blanka és Bacsó Zsuzsanna (*in*: Rend és Kosz [1997]) a magyar burgonyapiac viselkedésben figyelt meg kaotikus tartományokat. Mivel az egytermékes piacmodellt illesztették a megfigyelt adatokra, megvizsgálták a modell reakcióját az intervenciók változtatására: itt is megfigyelték a nem-lineáris dinamikai rendszerekben megszokott periodikus és kaotikus tartományok megjelenését, ill. periódus-többszörös bifurkációkat, a paraméterek változtatása során.

## 1.2 Biológia és statisztikus fizika

Érdekes a természettudományokra, nagyon nagy számban vannak a fraktális viselkedést mutató természeti jelenségek (Family & Vicsek [1991], Vicsek [1992]), a kristály-növekedéstől kezdve, a domborzat és a tengerpart vonalának alakulásig. Kaotikus jelenségek megfigyelhetők számos kémiai reakcióban (Szpálusy & Tl [1982]).<sup>6</sup>

Önszerveződő folyamatokra az élővilágból meríthetjük a legjobb példákat. Az egyik első vizsgált példa a *Dictyostelium discoideum* nevű amőbafaj sejtpopulációjának tövényszórási alakulása (Prigogine & Stengers [1986], pp. 148-149): ha a telep felemszította környezetben a tápanyagot, a sejtek szétválnak egy masszív, amelyben differenciális törvények egy szór, azon massa, és ebből sprók radnak ki, tápanyagban gazdag lóhelyek fel. Ez a folyamat "spontán módon" jön létre, az amőbapopulációban vándorló hullámok alakulnak ki, egy – a véletlen fluktuáció által létrehozott –

---

<sup>6</sup> A brüsszeli és a Bjeluszov–Zsobotyinszkij-reakcióval kapcsolatban ld. még Prigogine & Stengers [1986]-t (pp. 149-153), amely a következő cikkekre hivatkozik: A. Winfree: Rotating Chemical Reactions, *Scientific American*, vol. 230 (1974), 82-95.



*ciklikus AMP*-t (ciklikus adenzin-monofoszfot) kibocst vndorlsi kzpont fel, s ez a vndorls hozza ltre a komplex struktrt.

Hasonl komplex struktrk sponntn ltrejtett figyelte meg Vicsek et al. [1995] "nhajt" rszecskekból ll rendszerben: habr az egyes rszecskek, amelyek pldul egy madrraj tagjait modellezhetik, nllan mozognak, bizonyos kritikus sűrűsg fltt s zaj alatt fzistalakuls trtnik a rendszerben, a mozgásirnyok egysges rendbe rendeződnek. Ennek oka a kzeli szomszdek mozgásirnynak tlagbl definilt rvid tv klcsnhats, amely mgis az egzs rendszerre kiterjedő rendet, kollektv mozgst kpes ltrehozni.

Az evolci modellezsben is sikerrel lehetett alkalmazni s statisztikus fizika fogalmait (Geritz et al. [1997]). Ezek a megkzeltsek a klmbző evolcis stratgik elterjedst vagy kihalst vizsgáljk, ha j mutcik, azaz j evolcis stratgik jelennek meg. "Elgzsokat" figyeltek meg, azaz egy faj kettvltst kt fajj. A kutatk clja az, hogy a modell idealizltsgnak cskkentsvel az evolci egyre tbb jelensgt legyenek kpesek lerni.

Egzsges emberek kt szvverse kztt eltelt időtartam időbeli vltozsait vizsgálta H. E. Stanley et al. [1992], 24 rs skln. Ha  $B(n)$ -nel jelljk az  $n$ -ik s az  $n + 1$ -ik szvvers kztt eltelt időtartamot, a  $B(n)$  fggvny  $C(t)$  autokorrelcis fggvnye s  $S(f)$  Fourier-spektruma hatvnyfggvnyt kvető viselkedésből ( $C(t) \sim t^\gamma$ ,  $S(f) \sim f^\beta$ ) megllapthat, hogy ebben az idősorban hossz tv korrelcik vannak jelen (ezeket a fogalmakat rszletesen a 3.3 fejezetben trgyalom). Rvid tv korrelcik a lgzs temnek a szvversre kifejtett hatsval mg magyazhatk lennek, a hossz tv korrelcik ltre a szerzők nem adnak magyazatot.

A molekulris biologia fel haladva, egyre tbb sejt-szintű folyamat lersra is ajnlanak a fizikusok modelleket. Prigogine & Stengers [1986] az *ATP* (adenozin-trifoszfot) *ADP*-v (adenozin-difoszfot) trtnő lebontsban felfedezett osszcillik matematikai modellezst emlti (pp. 147-148.), amely a rekacisorozat kmiai kinetikjbl szksgszerűen kvetkezik. Msik plda erre a *kinezin ATP*-z viselkedésnek modellezse Deryi & Vicsek [1996]-ban.

Dolgozatom trgya, a DNS-szekvencik elemzse, a statisztikus fizika azon fejezeteihez tartozik, amelyek *szimblumsorozatok*, gy pldul természetes nyelven rott szvegek vagy szmt-gpes programok, statisztikai vizsgálathoz jrulnak hozz a fizika ltal motivlt tletekkel. Ezek kzl a mdszerek kzl ismertetek nhnyat a harmadik fejezetben, majd a negyedik rszben egy, az rott szvegekre mr bevlt, de a genetikban mg nem felhasznlt mdszer alkalmazst mutatom be DNS-szekvencikon.

## 2. fejezet

### A DNS-szekvencik kutatásnak irányai

#### 2.1 A statisztikai megközelítés kitérője

Az elmúlt években, másfelvélvélben exponenciálisan nőtt az ismert DNS-szekvencik mennyisége. Ma már szinte automatikussal automatizálható volt az a munka, ami vélvéldekkel ezelőtt még hosszú, vélvéltes és sok emberi lelemnyel ignylő feladatnak szmtott. A 2.1. táblázat és a 2.1. ábra az *NCBI-Genbank* (megtallható a <http://www.ncbi.nlm.nih.gov> címen) adatbázisban tárolt adatmennyiség időbeli vélvéltozást mutatja.

A szekvencik feldertésnek automatizálhatóságát vetette fel azt az igényt, hogy a szekvencik "jelentsenek" meghatározást is automatizálni kell ahhoz, hogy a kitérő munkafizetést tudjon tartani egymással. A problémát itt nem is a kódolási részek aminosav-szekvencivélvéltrnő lefordítása jelenti, hiszen a kódolási részek az 1960-as években megfejtették (Berend [1980], 41. o.). A feladat a kódolási és a nem kódolási részek kitérővélvélválasztása, a gének azonosítása, és ez a kódolási munka még messze van attól, hogy teljesen megoldottnak tekinthessék, különösen a magasabb rendű eukarióta lélvélnyek esetében (Fickett [1996]).

Fickett szerint a cél az "automatikus annotáció", azaz a szekvencia egyes szakaszaihoz hozzrendelni azok "tulajdonságait", minél teljesebb és pontosabb módon. Ezek a "tulajdonságok" sokféleképpen lehetnek, de a legjelentősebb kódolási az egyes szakaszok *funkcionális* szerepének (fehérje kódolása, szabályozás,...) az meghatározása. Ezen eredmények egyik nagyon fontos kitérője Fickett szerint – hosszú távon – olyan általánosságok levonása, amelyek szabályokat adhatnak a gének és genomok tervezéséhez, jövőbeli kutatások kezébe. Jelenleg viszont még meg kell elgondolnunk azzal, ha az algoritmusunk elfogadható pontossággal azonosítja meg, mely szekvencia kódolási fehérjét. Ha a fehérje funkciójára is kapunk tippet, gy az már jelentős eredmény. Az aminosav-szekvencit

Release	Datum	Bzisok száma	Ttelek száma
3	82. dec.	680338	606
14	83. nov.	2274029	2427
20	84. mj.	3002088	3665
26	84. nov.	3689752	4393
40	86. feb.	5925429	6642
44	86. aug.	8442357	8823
48	87. feb.	10961380	10913
53	87. szept.	15514776	14584
55	88. mrc.	19156002	17047
57	88. szept.	22019698	19044
60	89. jn.	31808784	26317
63	90. mrc.	40127752	33377
64	90. jn.	42495893	35100
67	91. mrc.	55169276	43903
69	91. szept.	71947426	55627
72	92. jn.	92160761	71280
73	92. szept.	101008486	78608
75	93. feb.	126212259	106684
79	93. okt.	157152442	143492
81	94. feb.	173261500	162946
83	94. jn.	191393939	182753
86	94. dec.	230485928	237775
90	95. aug.	353713490	492483
92	95. dec.	425860958	620765
94	96. pr.	499127741	744295
97	96. okt.	651972984	1021211
101	97. jn.	966993087	1491069
103	97. okt.	1160300687	1765847
105	98. feb.	1372368913	2042325

2.1. *tblzat: Az NCBI-Genbank adatbzisban trolt informci-mennyisg (ttelek száma, ill. az azokban szesen lvó bzisok száma) időbeli nvekedse. A GenBank-ben tallhat bzisok száma krlbell 14 hnaponknt ktszereződtt meg. (Forrs: <http://www.ncbi.nlm.nih.gov>)*

kzvetlenül nem kdol DNS-szekvencik automatikus csoportostsa, annotlsa mg nagyon tvolinnak tűnik, ugyanis, mint azt lttni fogjuk, ezek jelentős rsznek szerepről mg nincs sok elkpzelsnk. Ehhez kapcsolódik a msodik problmakr, amely a DNS-szekvencik kutatásnak irányt meghatározza.

A msodik nagy krds az, hogy mi a nem kdol szakaszok egy rsznek a jelentősge. Azon szakaszok, amelyek nem rdnek t kzvetlenül proteinn, sokféle szerepet tlthetnek be (Weaver

2.1. *bra*: Az NCBI-GenBank fejlődési téma, amely jól kezelhető exponenciálissal, s körülbelül 14 havonta kiegészül meg az ismert szekvenciák bizisainak összege. (Forrás: <http://www.ncbi.nlm.nih.gov>)

& Hedrick [1992]). Ezek egy része már már jól ismert. Példul a *TATA-box* 25 darab timin és adenin váltakozó sorozatból áll, és a legtöbb eukarióta sejt promotereinek jellegzetes része, amely az RNS-polimeráz bekapcsolását, és ezáltal a transzkripció megkezdését teszi lehetővé. Más szekvenciák szintén jelentős szerepet játszanak a transzkripció szabályozásában.

Vizsgálatok sok szekvencia szerepét még ismeretlen. Néhány kutató hajlanak arra, hogy ezek egy részét "evolúciós szemétként" tekintsék. Szerintük olyan szakaszokról van szó, amelyekről megszünt a transzkripció az evolúciós során, elvesztették jelentőségüket. Így szabadon, fenotipikus következmények nélkül, ezért könnyen mutálódhatnak, vagyis az eredeti "jelentséget" (példul kódolt aminosav-szekvencia, szabályozó szerep) már nem felismerhető.

Van olyan vélemény, amely szerint az intronok szerepe abban áll, hogy még *redundnsabb* tegyék a genetikai kódot, tovább csökkentve ezáltal a mutációk veszélyét. Kétségtelen, hogy már a kódszót is tartalmaz bizonyos redundanciát (a genetikai kód *degenerált*). Hiszen bizis-hármasok (triplettek, kodonok) kódolják az aminosavakat, márpedig a négy bizis 64-féle triplettet alkothat, és még ha le is vonjuk a *stop*-jelnek megfelelő három kodont, 61 bizis-hármas tartozik a 20 (21) aminosavhoz. Ugyanakkor, sokak szerint a nem kódolt szakaszok a proteinek elsődleges szerkezetét mutatják, biológiailag fontos információkat tartalmazhatnak (Mantegna et al. [1995b]).

Különböző "rejtélyes" az *intronok* szerepe. Ezek olyan szakaszok, amelyek az aminosav-szekvenciát kódoló *exonokkal* együtt terjednek a DNS-ről az mRNS-re (*primary transcript*, azaz elsődleges írást), viszont később valamilyen mechanizmus kivágja az intronokat az mRNS-ből, így azok nem kerülnek bele a proteint szintetizáló folyamatokba. Egy magasabb rendű faj *primary transcript*-je tipikusan tartalmaz egy *vezérszakaszt*, amelynek a szabályozásban van szerepe, egy *végszakaszt*, exonokat, valamint az exonokat egymástól elválasztó intronokat.

A génekben belfelől és a gének közötti nem kódolt szakaszok összetettebbé teszik a gént, genomokat. A 31 nukleotidból álló virális SV40 gént és a 210 000 bázist tartalmazó emberi dystrophin

gn kztt szrdnak az ismert gnek mretei (Mantegna et al. [1995b]). A genomok mrete is jelentős eltrseket mutat. Kt emltsre mlt paradoxon a kvetkező:

- *C-paradoxon* ("complete genome size" paradox): a genom komplexitsa nem ll semmi-  
fle korrelciban a fenotipikus komplexitssal. Pldaknt, a *Homo sapiens* genomjnak a mrete  
( $C \approx 3,4 \times 10^9 bp^1$ ) jval alatta marad a tdős-kopoltyis halnak ( $C \approx 1,4 \times 10^{11} bp$ ) vagy az  
*Amoeba dubia* nevű amőbafajnak.

- *Kdol s nem kdol rszek arnya*: A publiklt komplett genomok, kromoszmk s nagyon  
hossz DNS-szekvencik analizsból arra a kvetkeztetsre jutottak a kutatk, hogy a kdol s  
a nem kdol szakaszok arnya *cskken* az egyszerőbb fajoktl a magasabb rendű fajok fel.  
Pldaknt, a Mantegna et al. [1995b] cikkben elemzett emlős (ember, egr) szekvencik esetben  
5,3%, az lesztő kt kromoszma esetben 70% krli, mg az *E. coli* bakterium hat szekvencija  
s kt fg teljes genomja esetben a 80%-ot is meghaladja a kdol szakaszok arnya az egsz  
szekvenciohoz kpest. Ez a megfigyels magyazhat mind az "evolcis szemt"-hipotzissal, mind  
azzal az elkpzéssel, amely szerint ezeknek a szekvenciknak igenis van jelentőségk, pldul  
a transzkripci finomabb szablyozsban, s ezek a mechanizmusok fejlettebbek a magasabb  
rendű lőlnyekben. (Mantegna et al. [1995b])

## 2.2. A gnidifikci mdszerei

A feladat teht a DNS klnbző tpus szakaszainak, olyan "tulajdonskait" megtalálni, amelyek  
alapjn sztvlaszthatk ezek a tpusok. Egy-egy ilyen tulajdonsg nem csupn egy DNS-t annotl  
automatikus algoritmus alapjul szolgálhat, hanem hozzájruhat a klnbző szakaszok szerepnek  
mlyebb megtrshoz is. A kutatsok jelenlegi stidiumban elsősorban a kdol s a nem kdol  
szekvencik klnbzősgre, a *gnek identifklsra* kell szzpontostani az erőinket. Dolgozatomban  
tovbb szűktem a krdst, az exonok s az intronok eltrő viselkedst vizsglom.

A mdszereket *Borodovsky, Koonin s Rudd* alapjn,<sup>2</sup> Ficket [1996] kt csoportra osztja.  
Az első csoport, a "mintzat-kereső mdszerek" (*template methods, intrinsic approaches*), a  
keresett objektum "prototpus" rjk le, "tbb-kevsbe velős s elegns mdon". Ezek az eljrsok  
pldul egy promter elemről ismerhetik fel a kdol szekvencikat.

Azokat a DNS-szakaszokat, amelyek egy protein, mRNS vagy pre-mRNS bektődst  
teszik lehetőv, s ezltal kulcsszerepet jtsszanak a transzkripciban, "signal", "ktőhely" (*bind-*

---

<sup>1</sup> A *bp* a *bzispr* (*base pair*) kifejezs rvidtse, s a genetikai irodalomban a DNS-szekvencik  
hossznak mrtkegsgeként használjk. Hasznlatosak a *bp* tbszrsei, a *kbp* s a *Mbp* is.

<sup>2</sup> M. Borodovsky, E. V. Koonin, K. E. Rudd: New genes in old sequence: a strategy  
for finding genes in the bacterial genome, Trends Biochem. Sci. **19**, 309-313.

ing site), "felismerő hely" (*recognition site*), vagy "sequent element" nven emlti a szakirodalom. A "mintzat-kereső mdszerek" legelterjedtebb vlfaja a "signal-keress" (*gene search by signal*), amely valamely, konkrét funkció betöltő szekvenciát keres a genomban. De mivel egy "signal"-fajtnak is sok változata létezik, először "konszenzus-szekvencia" definícióval próbkoztak: egy adott "signal" összes előfordulást egymás mellé helyezték, s az idealizált szekvenciát így kapták meg, hogy mindegyik pozícióba az ott leggyakrabban előforduló bázist vették fel.

De ez a módszer túlságosan egyszerűnek s nem elég hatósónak bizonyult, ezért az *n*. *matrix-módszerhez* kellett folyamodni. Jelleme  $f(b, i)$  a  $b$  bázis előfordulási gyakoriságát a  $i$ -edik szekvenciájában,  $p(b)$  pedig a  $b$  bázis frekvenciáját a genomban. Az

$$m(b, i) := \log \left( \frac{f(b, i)}{p(b)} \right) \quad (2.1)$$

matrixot nevezzük a "signal" "pozíció-súly matrixnak" (*position weight matrix*). Amikor egy szekvenciáról el akarjuk dönteni, vajon az "signal"-szekvencia-e, képeznünk kell a tesztelendő szekvencia pozíció-súly matrixát is:

$$s(b, i) := \begin{cases} 1, & \text{ha a szekvencia } i\text{-ik pozíciójában a } b \text{ bázis szerepel} \\ 0 & \text{egyébként} \end{cases} \quad (2.2)$$

A teszt eredményt, amely megadja, vajon mennyire "hasonlít" a tesztelt szekvencia a "signal"-ra, a két matrix skalárszorzataként kapjuk meg:

$$S := \sum_{b,i} [m(b, i) \cdot s(b, i)]. \quad (2.3)$$

Ha ez az  $S$  érték magas, nagy valószínűséggel lesz a vizsgált szakasz "signal". Az eljárásért továbbfejlesztése már a géneket felismerő mdszerek második osztályának technikai részleteihez illik.

A "mintzat-kereső mdszerekkel" szemben llnak, s utóbbiak egyre nagyobb szerephez jutnak az ismert szekvenciák számának rohamos növekedésével, a "visszakereső mdszerek" (*lookup* vagy *extrinsic methods*). Ezen algoritmusok az ismert szekvenciák összehasonlításából vonnak le következtetéseket. A legjobb példát ez utóbbi eljárás csoportra a szekvenciák közötti hasonlóságot definiáló mdszerek adják.

Két, azonos hosszúságú DNS-szekvencia közötti "távolság" legegyszerűbb mértéke így határozható meg, hogy pontosan összehasonlítjuk a két szekvenciát alkotó bázisokat, s vesszük az eltrő bázisok számát (Weir [1990]). De ezen módszer, illetve továbbfejlesztett változatai, csupán bizonyos esetekre, magához a biológusok számára igen fontos, fajok közötti genetikai távolságok meghatározására

alkalmasak, amelyek az evolúciós-genetikai rokonságok meghatározásához szükségesek (*phylogeny construction, parsimony methods*).

Ennél finomabb, szélesebb körben használható módszerek a szekvenciák bizonyos statisztikai tulajdonságait használják fel. Fickett [1996] közel száz ilyen módszert sorol fel, a kodonok, hexamerek (bizonyosok), aminosavak,<sup>3</sup> aminosavpárok előfordulási gyakoriságát kódoló vektorokkal kezdve, a szekvencia Fourier-transzformáltjait tartalmazó vektorig. Ezekkel a módszerekkel a géneknek legalább a feltételezhetően már ma is automatikusan felismerési, és sok esetben a gén funkcióira is lehet következtetni.

Ezek a módszerek a gének azonosítására szolgálnak. A dolgozatomban fő irányom az exonok és az intronok, ill. az intronokat tartalmazó és nem tartalmazó gének közötti statisztikai különbségek vizsgálata volt. Ezen különbségek felismerése egyszerre járulhat hozzá mindkettő, a 2. fejezet elején felvetett probléma megoldásához: egyszerűen alapjaiban változhatnak olyan javított automatikus algoritmusoknak, amelyek képesek lesznek szétválasztani a kettő szekvenciát, másrészt az intronok tulajdonságainak megismerése nyitva tehető a jövőben az intronok szerepére is.

---

<sup>3</sup> Az aminosavak frekvenciája Mantegna et al. [1995b] szerint csupán kis különbségeket mutat, ha nagyobb szekvencia-halmazokra tagolva vizsgáljuk. Ezzel szemben, a kodonok gyakoriságának eloszlásában nagy eltéréseket találhatunk.

### 3. fejezet

## DNS-szekvencik s rott szvegek modellezse statisztikai eszkzkkal

Szimblumszekvencik statisztikai megkzeltsei kzl a legegyszerűbb az egyes szimblumok előfordulsi gyakorisgainak (frekvenciinak) az elemzse. Kzismert pldul, hogy egyszerűbb titkosrsok megfejtsnl az egyik leggyakrabban hasznlt "kezdőlps" a legtbbbszr előfordul jel megkeresse szokott lenni, s ez a szimblum ltalban a leggyakoribb betűnek, az 'e'-nek felel meg. Nagy Ferenc [1986] sok tblzatot hoz a magyar s ms nyelvek betűinek, hangjainak az eloszlsrl. Megvizsgltk az eloszlst kln szkezdő, szkzepe s szvgi helyzetben is.

Weir [1990] az egyes bzsok gyakorisgra hoz pldkat, klnbző lokuszokban.<sup>1</sup> Pldaknt, Weir 7.5 tblzatban (233. oldal) szerepel a karfiol mozaik-vrus, amely MCACGDH gnjben az adenin arnya 37%, mg a  $\Phi$ X174 bakteriofg PX1CG lokuszban csupn 24%. Az egr mitokondriumban 12% a timin arnya, a Hepatitis B vrus HPBAYW gnjben viszont 27%.

A biológusok a leggyakrabban a DNS-szekvencik n. *GC-arnyt* (vagy *G+C-arnyt*) vizsgáljk. Egy DNS-szekvencia *GC-szintje* a guanin s a citozin molris koncentrcijt jelenti az adott gnben, molekulban vagy genomban, amely rtk jellegzetesen 30% s 80% kz szokott esni.

*Bernardi* s csoportja az 1980-as vektől kezdődően vizsgálja klnfle szervezetek genomjainak szerkezetét, a GC-szint fggvnyben. Salinas et al. [1988] kimutatja, hogy egyes nvnyfafajok genomja mozaikszerűen pl fel a GC-szint szempontjbl homogén sszettelű alkotelemekből. Egy-egy ilyen elem 100 – 200kbp-nl is hosszabb lehet, s Matassi et al. [1989] ezen elemeket "izokroknak" (*isochores*) nevezi. Klnbző nvnycsaldok eltrő izokr-mintzattal rendelkeznek. Hasonl izokr-szerkezetet mutattak ki a gerincesek genomjban is (Montero et al. [1990]).

---

<sup>1</sup> A lokusz (*locus*) olyan gnt jell, amelynek ismert a pontos pozcija a kromoszmn.



Ezek a kutatások egyttal az exonok s az intronok GC-szintjvel kapcsolatban is hoztak új eredményeket. A Salinas et al. [1988] által vizsgált növényekben a GC-szint jóval alacsonyabb valamely gén intronjaiban, mint az adott gén exonjaiban, de a két érték között lineáris korrelci fedezhető fel egyes fajok genomján belül. Az is megfigyelhető, hogy azoknál a fajoknál, amelyeknél tagosan alacsonyabb az exonok GC-szintje, alacsonyabb az intronok GC-szintje is. Hasonló eredmények szerepelnek, szintén növények kapcsán, Carels et al. [1998] 6. részén, valamint hasonló korrelciokról szólnak be az emberi genomban Clay et al. [1996]. Ezek a tények részben indokolhatják a 4. fejezetben ismertsre kerülő saját eredményeimet is.

A kódoló szekvenciákban, a triplettek (kodonok) eltrő pozícióiban eltrő az egyes bázisok gyakorisága. Herzel & Große [1995], ill. Herzel et al. [1997a] az emberi III. kromoszómájának adatait, Herzel & Große [1997b] pedig az emberi  $\beta$ -miozin HUMBYH7 gén ritkkeit idézi: valamely bázisnak a kodon első, második, ill. harmadik pozícióban való előfordulási gyakorisága között gyakran két-háromszoros szorzás is lehet.<sup>2</sup>

A különböző pozícióban mérhető GC-szintekre  $GC_1$ ,  $GC_2$ , ill.  $GC_3$ -ként szokás hivatkozni. Mivel a harmadik pozíció általában redundáns (legalább részben<sup>3</sup>), ezért releváns a  $GC_{1+2}$ -szint is, amely a guanin s a citozin koncentrációját jelenti az első két pozícióban. A különböző pozícióban mérhető GC-szintek közötti lineáris korrelciokról szintén beszólnak Bernardi csoportjának idézett cikkei.

Továbbá az egyes bázisok frekvenciáján, vizsgálhatjuk nukleotid-kettesek gyakoriságát is. Weir [1990] erre is hoz példát (ld. Weir 7.8. táblázat). Megfigyelhető, hogy a szomszédos bázisok nem függetlenek egymástól, vagyis ha  $p_i$  jelöli az  $i$  bázis előfordulási valószínűsége,  $p_{ij}$  pedig az  $(ij)$  bázis-kettes gyakoriságát, akkor

$$p_{ij} \neq p_i p_j. \quad (3.1)$$

Erre szolgáltat példát Weir [1990] 7.7 táblázata.

### 3.1 Markov-folyamatok

A szomszédos bázisok közötti korrelciót szolgáltatva az ötletet, hogy a DNS-szekvenciákat *Markov-láncként* lehetne modellezni (Weir [1990]). Az egyszerű statisztikákon támaszkodva, a sztochasztikus

---

<sup>2</sup> A legszűkösebb különbség a HUMBYH7 gén második s harmadik kodon-pozíciója között fordul elő: az adenin frekvenciája a második pozícióban 43,7%, míg a harmadik pozícióban csupán 7,9%!

<sup>3</sup> Példul mind a nagy TC-vel kezdődő kodon a *szeryin* nevű aminosav megfelelője a kódszótárban; a CAT s a CAC *hisztidint* kódoló, a CAG s a CAA pedig *glutamint* (Berend [1980]).

eszkzkkal trtnő modellezs j skra emeli a DNS-szekvencik tanulmnyozst, s ezen a ponton tallkozik a genetika a statisztikus fizika 1. fejezetben említett interdiszciplnrns alkalmazsaival.

Sztochasztikus folyamatnak egy olyan fizikai (vagy bármilyen egyéb valóságos) rendszert vagy matematikai modellt nevezünk, amely "egy valószínűségi sorozat által szabályozott szimbólumsorozatot produkál" (Shannon & Weaver [1986], p. 55). Ezzel a jelzővel a nem determinisztikus, csupán statisztikai eszközökkel elemezhető idősorokat szokás jelölni, legyen sz tőzsdei rfolyamok alakulsrl, vagy a szvversek kztt eltelt időintervallumokrl (pldkat ld. az 1. fejezetben). De a sztochasztikus folyamatok prototípusai kezdettől fogva a karakter szekvenciák, például egy természetes nyelven írott szöveg szimbólumainak sorozata. A szimbólumszekvenciák hagyományos sztochasztikus elemzési eszközei a *Markov-láncok* és *Markov-modellek*. rthető teht, hogy ezekkel prblkoztak először (az 1980-as vekben) a genetikusok is.

Egy  $n$  elemű ábécé (szimblumhalmaz) feletti *Markov-láncot* egy  $P$  mátrix definiál, ahol a  $p_{ij}$  mátrixelem ( $i, j = 1..n$ ) annak a feltételes valószínűségét jelenti, hogy a "mondat" valamely pozíciójában az ábécé  $j$ -ik betűjét találjuk, feltéve, hogy a megelőző pozícióban az  $i$ -ik betűt találjuk. Az ún. *ergodikus Markov-láncok* esetén (a lánc elejétől elég távol) az egyes szimblumok előfordulási valószínűsége független a pozíciótól.

Magasabb rendű ( $n$ -ed rendű) Markov-láncot úgy kapunk, ha a következő karakter valószínűsége a megelőző  $n$  betűtől függ. (Tehát a fentebb lert Markov-lánc az  $n = 1$  esetnek felel meg, mg a 0-ad rendű Markov-lnc fggetlen szimblumok sorozata.)

A Markov-folyamatoknak egy ltalnosabb osztlyt jelentik a *Markov-modellek*. Egy *Markov-modell* esetén, adva van egy véges állapothalmaz ( $\Omega = \{s_1, \dots, s_n\}$ ), egy véges ábécé ( $\Sigma = \{\sigma_1, \dots, \sigma_m\}$ ), egy  $n \times n$ -es  $P$  állapotátmeneti mátrix és egy  $n \times m$ -es  $A$  jelkibocsátási mátrix. A modell úgy "működik", hogy ha a modell az  $i$ -ik állapotban van,  $a_{ik}$  valószínűséggel kibocsátja a  $k$ -ik szimblumot, majd  $p_{ij}$  valószínűséggel átmegy a  $j$ -ik állapotba, és újból jelet bocsát ki, stb.

Markov-lánc esetén, a modell paramétereit, azaz a  $P$  mátrix elemeit megbecsülhetjük kellően nagy korpusz feldolgozásával, azaz a megfelelő karakter- és karakterpár-gyakoriságok empirikus megfigyelésével. Markov-modellek esetén viszont, általában nem ismert az, hogy adott korpusz mögött milyen, azt létrehozó állapotsorozat húzódik meg. Ez esetben meg kell találnunk, mely  $S \in \Omega^*$  állapotsorozat esetén maximális a  $P(S|O)$  feltételes valószínűség, ahol  $O \in \Sigma^*$  a megfigyelt jelsorozat. (Az  $\Omega^*$  s a  $\Sigma^*$  az  $\Omega$ , ill. a  $\Sigma$  elemeiből kpezett vges hosszsg sztringek halmaza.) Azaz, Bayes törvénye miatt, maximalizálni kell a  $P(O|S) \cdot P(S)$  szorzatot. E célra különböző iteratív algoritmusokat dolgoztak ki (Viterbi-algoritmus, stb.), melyek segítségével meg lehet becsülni a Markov-modell paramétereit

(”rejtett Markov-modellek”; Rabiner [1989], Krenn & Samuelsson [1996]).

A DNS modellezése Markov-folyamatként Weir [1990] szerint inkább az egész genom, mint egyes gének szintjén szokás, mivel az egyes gének hossza a legtöbbször nem elég hosszú ahhoz, hogy a Markov-lépcső rendjét meg lehessen határozni. Így például egy 1000bp hosszúságú szekvencia esetén csupán a másodrendű Markov-jelleget lehetne kimutatni, az viszont valószínűtlen, hogy bizonyos szisztematikus frekvenciák kellően lerázza a gént. (Megjegyzem, hogy az a jóval hosszabb gének is ismertek, s ugyanakkor a Markov-lépcsővel történő modellezés elvesztette az érdekességet.) A modell alkalmas arra, hogy egyes részeket, bizonyos  $n$ -esek gyakoriságát megjósoljuk. Példul, ha kiszámoljuk egy restriktív szekvencia<sup>4</sup> gyakoriságát a genomban, megmondhatjuk, hány kromoszomadarab keletkezik, amikor egy adott restriktív enzim működésbe lép a genomon. Becslést adhatunk arra is, vajon milyen hosszú lesz a leghosszabb ismétlődő szakasz valamely szekvenciában.

A Markov-folyamatok ugyanakkor nem elégségesek se nem a természetes nyelvi, se nem a genetikai szekvenciák modellezésére. Chomsky [1957] a *Syntactic Structures*-ben amellett érvel – és a generatív szintaxis azóta is alátámasztotta ezt a gondolatmenetet –, hogy az angol, és általában a természetes nyelvek szintaxisa nem írható le Markov-folyamatokkal, ill. reguláris grammatikával, csupán környezetfüggetlennel (ld. a 3.6.1. fejezetben). A következők fejezetben olyan eredményeket ismertetek, amelyek a Markov-modellek elgátlensége mellett hoznak fel bizonyítékokat a DNS-szekvenciák esetében is.

## 3.2 A Zipf-analízis

G. K. Zipf már az 1930-as években végzett szisztematikus vizsgálatokat természetes nyelven írott szövegekkel (Zipf [1935, 1949]). Szisztematizálta nagy korpuszokban az egyes szavak előfordulási gyakoriságát, s sorrendbe helyezte az egyes szavakat gyakoriságuk szerint:  $R = 1$  jelentette a leggyakoribbat,  $R = 2$  a második leggyakoribbat, s így tovább. Ezután bebiztosította az  $\omega$  gyakoriságot (frekvenciát) az  $R$  sorrend függvényében. Az eredmény meglepő volt: az  $\omega(R)$  függvény nem exponenciálisan csengett le, ahogy vártunk, hanem hatványfüggvényszerűen:

$$\omega(R) \sim R^{-\zeta}, \quad (3.2)$$

ahol a  $\zeta$  kitevő, nyelvtől és a korpusz tartalmától függetlenül, univerzálisan 1 körül értéket vett fel:  $\zeta \approx 1$ . Ez azt jelenti, hogy log-log skálán ábrázolva, az  $\omega(R)$  függvény egy  $-\zeta$  meredekségű egyenest ad. A Czirk et al. [1995] által idézett irodalom szerint, a Zipf-analízist elvégzők városok, valamint ipari vállalatok méret szerinti eloszlása vizsgálatra is.

---

<sup>4</sup> Olyan szekvencia, amelyhez egy restriktív enzim kapcsolódhat, s így elvághatja a szekvencia közepén a DNS-szálat.

A módszer alkalmazsa DNS-szekvencik esetén felveti azt a kérdést, hogy vajon mit tekinthetünk "sznak" a genomban? A kódol szekvencik esetén könnyen addik a válasz: a triplettek (kodonok) a legkisebb "értelmes" ptőelemek. Viszont a nem kódol szakaszok jelentős résznek funkciója még nem ismert, ezért Mantegna et al. [1994, 1995b] bevezeti a Zipf-analízis egy szerű általánosított, amelyet magyarul "szimblum  $n$ -es Zipf-analízisnek" nevezhetünk (*n-tuple Zipf-analysis*). Ennek lényege az, hogy egy  $n$  szimblum hosszúság "ablakot" mozgatunk végig a szekvencián, minden lépésben egyetlen (tehát nem  $n$ ) karakterrel elmozdítva az ablakot, és az így kapott jel  $n$ -esek frekvenciára szüntetjük ki az  $\omega(R)$  gyakoriság-sorrendben elfoglalt hely függvényét.

A szimblum- $n$ -es Zipf-analízis természetes nyelvi szövegek esetében ppgy hatványfüggvény eredményez, mint a hagyományos Zipf-analízis, viszont eltérő lesz az exponens. Mantegna et al. [1995b] egy angol nyelvű enciklopédia cikkeit elemezték ilyen eszközzel: a hagyományos, szavak gyakorisága alapján végzett Zipf-analízis  $-0,85$  kitevőt eredményezett, míg a betű- $n$ -esek ( $n = 3, 4, 5, 6$ )  $-0,57$  körüli meredekségekkel jellemezhető brkat produkáltak a  $10 \leq R \leq 1000$  közötti nagyságrendet tfog intervallumban. A két exponens különbségt a szerzők abban látják, hogy még a hagyományos Zipf-analízisben csupán egy meghatározott szókincs elemei vesznek részt, addig az  $n$ -esekkel végzett Zipf-analízis során kibővíti ez a halmaz.

Kipróbálták ezt a szimblum- $n$ -es Zipf-analízist mesterséges nyelvekre, mghozz a UNIX operációs rendszer kompilált fájljaira is ( $n = 6, 8, 10, 12$ ). A 9000000 bit hosszúságú szöveg egy kétélemtű bcből ( $\{0; 1\}$ ) pl fel, és a DNS-adatokkal való könnyű összevetethetőség érdekében (új elemű bcb) választották meg így az  $n$ -eket. Ebben az esetben is lineáris brkat kaptak a kétszeres logaritmikus skálán, és a lineáris tartomány szívehetően nőtt  $n$  növeléssel. Az  $n = 12$  mellett  $\zeta = 0,77$  exponenst kaptak a szerzők.

Mantegna et al. [1994] és Mantegna et al. [1995b] az akkoriban hozzáférhető leghosszabb emlős, gerinctelen és vírus DNS-szekvencikra alkalmazta a Zipf-analízist. (Prokarióta szervezetek, illetve mitokondriumból és kloroplasztiszból származó szekvencikkal nem foglalkoztak.) Ahhoz, hogy egy  $L$  hosszúságú szakaszt  $n$ -es Zipf-analízisnek lehessen alvetni,  $L \gg 4^n$ -nek teljesnie kell (a cikkben  $L > 10 \cdot 4^n$  teljesül). A cikkben az  $50000bp$ -nél hosszabb, vagy az ezt az értéket megkzélő hosszúságú szekvenciket használták fel, így az  $n = 3, 4, \dots, 7$ -eseket lehetett csak vizsgálni, de a különböző  $n$ -ekre nem kaptak lényegesen eltérő eredményeket. Nyitott kérdés, hogy vajon a  $\zeta$  értéke konstans vagy monoton nő, ha növeljük  $n$  értéket. E kérdés megválaszolásához ( $\zeta$  skálási tulajdonságainak feldertéshez) az 1995-ben hozzáférhető szekvenciáknál jóval hosszabb szekvencikra lett volna szükség. Fontos még megjegyezni, hogy az adatbázis felhasznált ttelei messze nem reprezentatvak az élővilág egészére nézve: a GenBank adatai között nagyobb arányban szerepelnek a kódol szakaszok, mint az élővilágban általában, egyes fajok, csoportok adatai többségben vannak más csoportokkal szemben, továbbá gyakran szerepelnek az adatbázisban különböző

fajok rokon szerepet betöltő szekvenciái is.

Ezen hosszú DNS-szekvenciák Zipf-analízise a Zipf-törvényben megszokott hatványfüggvényt eredményezett. Példul, az emberi HUMRETBLAS szekvencia  $\zeta = 0,33; 0,34; 0,34$  exponenst adott rendre  $n = 4; 5$  s  $6$  mellett, az  $R \leq 4^{n-1}$  tartományban. Az  $R \geq 4^{n-1}$  régióban gyorsan lecsökken a frekvencia-függvény, mivel néhány basis  $n$ -es nem fordul elő a 180 000 karakterből álló szekvenciában.

A következő lépésben a szerzők az egyes szekvenciákat kódként jelsorozattal bontották szét: a GenBank-beli fajl információi alapján kiválasztották az egyes szekvenciák ismert kódjait a többi szakasztól. A kódok s a nem kódok<sup>5</sup> genom-elemek konkatenciái (szszefűzése) erősen eltérő Zipf-viselkedést mutattak. A nem kód szekvenciák a természetes nyelvi szvegekhez hasonló, (3.2) alak hatványfüggvényt eredményeztek az  $R \leq 4^{n-1}$  intervallumban. Ezzel szemben, a kód szakaszok Zipf-analízise az  $R \leq 4^{n-1}$  tartományban *logaritmikus* függvénnyel közelíthető gráffal adott, amely a lin-log skálán lineáris:

$$\omega = b - c \log_{10} R. \quad (3.3)$$

Ezt az eredményt Mantegna et al. [1995b] kellő számú és kellő hosszúságú szekvenciára kimutatta,  $n = 3, 4, \dots, 7$ -re. Remélem megjegyezni, hogy ez a cikk és Kanter & Kessler [1994] szerint (3.3)-hoz hasonló eloszlást követnek az abc betűinek gyakoriságai az emberi nyelveken és szvegekben is.

Mit jelent ez a megfigyelés? Czirk et al. [1995, 1996] vizsgálja azt a kérdést, hogy milyen típusú "szvegek" milyen Zipf-függvényt eredményeznek. Levezeti, majd szimulációval is igazolja azt, hogy a Markov-modellek Zipf-függvényei lépcsőzetesen futnak le, szemben a különböző módszerekkel előállított, hosszú tv korrelációt mutató szekvenciák hatványfüggvény szerinti lecsökkenésével. Utóbbiak jól közelíthetők Markov-modellekkel a központi tartományban, de a Zipf-plot kódként szelvényesen eltér az illesztett Markov-modellből származó gráffal. (Működő magyarázatot javasol a Zipf-törvény és a Markov-folyamatok összekapcsolására Kanter & Kessler [1994].)

Vajon illeszthető-e Markov-modell a DNS-szekvenciákból származó  $\omega(R)$  függvényekre? Mantegna et al. [1995b] szerint az egyes szekvenciákból empirikusan származhat  $p_{\sigma_1 \sigma_2}$  átmeneti mátrix és  $v_\sigma$  gyakorisági vektor elemei alapján

$$P(\sigma_1 \sigma_2 \dots \sigma_n) = v_{\sigma_1} p_{\sigma_1 \sigma_2} \dots p_{\sigma_{n-1} \sigma_n} \quad (3.4)$$

---

<sup>5</sup> A "nem kód szakasz" kifejezés itt, az eddigiekben és az elkövetkezőkben is, sokszor azt jelenti, hogy az adott szakasz valódi szerepe nem ismerjük, vagyis nincs kizárva az sem, hogy egyes részeit proteint kódoljanak. De valószínűleg más a szerepe, ha van egyáltalán.

adja meg a  $(\sigma_1\sigma_2\dots\sigma_n)$  bzis  $n$ -es gyakorisgt az illesztett első rendű Markov-lncban (ahol  $\sigma_i = \text{A, C, G, T}$ ). Ez a prblkozás a kdol szakaszok esetn elg j kzeltst ad. Viszont a nem kdol rszek (ill. az intronban gazdag tartományok) esetn az első msfl nagysgrendben szignifikns az eltrs a megfigyels s a modell kztt.

A tovbbi ksrletek azt mutatjk, hogy a Markov-lnc rendjnek nvvelvel egyre jobban lehet "utnozni" a megfigyelt statisztikai tulajdonsgokat. Mgis, sszhangban a kvetkezőkben ismertetsre kerlő eredmnyekkel is, le kell vonnunk azt a kvetkeztetst, hogy a *markovi sztochasztikus folyamatok nem rjk le elg jl a nem kdol rszeket*.

### 3.3 Hossz tv korrelcik

Az, hogy a DNS-szekvencik nem modellezhetők Markov-folyamatok segtsgvel, mr a Zipf-analízis alkalmazsa előtt is ismert volt: Stanley s trsai kutatsai (Peng et al. [1992], Stanley et al. [1992, 1993a, 1993b]) hossz tv korrelcik ltt mutattk ki egyes DNS-szekvencikban. Mrpedig Markov-folyamatokban (leszmtva a determinisztikus Markov-folyamatokat, amelyek egy specilis, nem jellemző hatresetet jelentenek) nem jelenhetnek meg hossz tv korrelcik, csupn rvid tvak, amelyek egy karakterisztikus  $R$  tvolsg alatt lecsengenek.

*Korrelcin* azt a jelensget rtjk, amikor egy  $X$  s egy  $Y$  mennyiség vltözsa statisztikailag "sszefgg": ha az  $X$  adatsor  $i$ -ik eleme,  $X_i$  nagyobb, mint az  $X$ -ek  $\langle X \rangle$  tlag, akkor "ltalban"  $Y_i$  is nagyobb, mint az  $\langle Y \rangle$  tlag. Egzaktabb formban:

$$\langle X \cdot Y \rangle - \langle X \rangle \langle Y \rangle > 0. \quad (3.5)$$

*Antikorrelcin* azt a jelensget rtjk, ha  $X_i$  tlagot meghalad rtke  $Y_i$  tlag alatti rtkvel jr tipikusan egytt, s fordttva:

$$\langle X \cdot Y \rangle - \langle X \rangle \langle Y \rangle < 0. \quad (3.6)$$

Egy "hossz"  $X_i$  szimblumsorozat esetn beszélhetnk *autokorrelcirl* is: ekkor a kt adatsor szerept  $X_i$  s annak  $k$  szimblummal val "elcssztatottja",  $X_{i+k}$  veszi t, azaz  $X_i$  s  $X_{i+k}$  vltözsa korrellhat, klnbző  $i$ -kre. A

$$C(k) := \langle X_i \cdot X_{i+k} \rangle - \langle X_i \rangle \langle X_{i+k} \rangle \quad (3.7)$$

fgvnyt, ahol az tlagolsokat az  $i$  pozcokra vgezzk el, az adatsor, ill. szimblumszekvencia *autokorrelcis fggvnynek* nevezzk. Amennyiben korrellatlan az adatsor (pldul egy kockadobsok eredmnyeiből ll sorozat esetn), gy  $C(k) = 0$ ,  $k \neq 0$ -ra. *Rvid tv korrelcikrl* akkor beszélhetnk,

ha létezik olyan  $R$  karakterisztikus tvolsg, amely szerint  $C(k)$  exponenciálisan lecseng (példul Markov-folyamatok esetén):

$$C(k) = C_0 e^{-k/R}. \quad (3.8)$$

Amennyiben nem létezik ilyen  $R$  karakterisztikus tvolsg, hanem  $C(k)$  hatványfüggvény szerint cseng le,

$$C(k) = C_0 k^{-\gamma}, \quad (3.9)$$

akkor *hossz tv korrelci* van jelen a szimblumszekvenciban. Kimutatható, hogy  $n$ -ed rendű Markov-folyamatok nem produkálhatnak hossz tv korrelciakat, azok lecsengenek  $n$  nagyságrendjében.

Hossz tv korrelciakat több módon is kimutathatjuk valamely szekvenciben. Az első megoldás a  $C(k)$  függvény felvétele közvetlenül, az adatokból. Ennek viszont meglehetősen nagy a számítási igénye. Viszont az alábbiakban bemutatandó "vletlen bolyongs-mdszer" egy olyan mennyiséget vezet be, amely szintén kimutatja a hossz tv korrelciakat, viszont diszkrét Fourier-analízissel egy "csúsztatott doboz" (*sliding box*) módszerrel könnyen számíthat.

A *vletlen bolyongs-modell*t először 1992-ben C.-K. Peng és társai, többek között a Bostoni Egyetem munkatársai (Peng et al. [1992], Stanley et al. [1992, 1993a,b], Peng et al. [1993]) alkalmazták DNS-szekvencikra. Később más fizikusok felhasználták a módszert írott szövegek elemzésére is (Schenkel et al. [1993], Amit et al. [1994], Dietler & Zhang [1994], Ebeling & Neiman [1995]).

Az eljárás lényege a következő: kiegészítjük elől egy bolhát, amelyet egy szimmetrikus origóba helyezünk. Felolvassuk neki a szekvenciát, és amennyiben pirimidinvesz bázisot talál (C vagy T), úgy előre ugrik egyet, míg purinvesz bázis esetén (A vagy G) hátra ugrik. Más szabályok is elképzelhetők, például A esetén (ill. C esetén, G esetén, T esetén) lép csak előre a bolha, míg az összes többi bázis esetén hátra; vagy erős hidrogénkötés esetén (C vagy G) lépnek előre, gyenge hidrogénkötés esetén pedig hátra; vagy az n. hibrid (vagy alfabetikus) szabály esetén: A és C jelenti az előreugrást, G és T pedig a hátraugrást. De a "purin-pirimidin"-szabály alkalmazása a legelterjedtebb, a következőkben is erre fogok hivatkozni. Elképzelhető persze az is, hogy valaki az egyes bázisokhoz nem egész nagyságugrást rendel (pl. a molekuláris tömeget vagy a hidrofobicitást javasolja a szakirodalom).

Amennyiben  $u_i$ -vel jelöljük az  $i$ -ik lépést ( $u_i = \pm 1$ ), úgy a bolha helyzete az  $i$ -ik lépés után nyilván

$$y(l) = \sum_{i=1}^l u_i. \quad (3.10)$$

Az  $y(l)$  mennyiség vrhat rtke kevs informcit nyjt, hiszen azt az egyes bzisok relatv gyakorisga hatrozza meg. Viszont annl rdekesebb mennyisg az  $y(l)$ -nek az elmozduls tlagakrl vett  $F(l)$  tlagos fluktucija (*root mean square fluctuation*):

$$F^2(l) := \left\langle (\Delta y(l) - \langle \Delta y(l) \rangle)^2 \right\rangle = \langle \Delta y(l)^2 \rangle - \langle \Delta y(l) \rangle^2, \quad (3.11)$$

ahol

$$\Delta y(l) = y(l_0 + l) - y(l_0), \quad (3.12)$$

s az sszes lehetsges  $l_0$  pozícira kell az tlagolst kpezni.

Az  $F^2(l)$  szoros sszefggsben ll a (3.7)-ben definilt  $C(k)$  autokorrelcis fggvnyvel:

$$F^2(l) = \sum_{i=1}^l \sum_{j=1}^l C(j - i). \quad (3.13)$$

Ebből kvetkezik, hogy  $-C(k)$  viselkedstől fggően  $-F(l)$  hromfle viselkedst mutathat. Mindhrom esetben  $F(l)$  hatvnyggvny-szerű viselkedst kvet,

$$F(l) \sim l^{-\alpha}, \quad (3.14)$$

s az eltrs az  $\alpha$ -exponensben nyilvánul meg:

- Amennyiben tisztán vletlen sorozattal llunk szemben, azaz  $C(k) = 0$  tlagban (kivve a  $k = 0$  esetet, hiszen  $C(0) = 1$ ), akkkor klasszikus, korrelatlan bolyongsrl van sz:  $\alpha = 0, 5$ .

- Rvid tv korrelcik eseten az autokorrelcis fggvny (3.8)-szerinti exponencilis lecsengse miatt az  $F(l)$  – egy "tranzien" utn – aszimpttikusan szintn  $\alpha = 0, 5$  kitevőjű hatvnyfgvnyt kvet.

- Hossz tv korrelcik eseten viszont,  $C(l)$  nem jellemezhető  $R$  karakterisztikus hosszal, (3.9) szerint viselkedik, s ezért  $F(l)$   $\alpha$ -kitevőjnek az rtke eltr a  $0, 5$  rtktől. Az eltrs mrtke jellemzi a hossztv korrelcik erőssgt. Az  $\alpha = 1$  eset felel meg a sklainvarins  $1/f$  zajnak ("maximal complexity limit") (Amit et al. [1993]). Az albbi esetekben  $0, 5$  s  $1$  kztti rtkekkel fogunk tallkozni.



Az eljárás tvihető természetes nyelven írott szvegekre is. Ebben az esetben egy-az-egyhez kódot alkalmaztak, azaz titk a szveget egy bináris bit segítségével (szemben a "DNA-walk"-kal, ahol is kódot a leggyakrabban ugyanígy kódolták). Példul a biten kitűnően lehet kódolni az angol bit 26 betűt, a szöveget, s a legfontosabb írásjeleket.

Az eredmények nagyon érdekesek. A vizsgált DNS-szekvenciák nagyon jól szétválasztható kódot csoportba (Peng et al. [1992]): a tisztán kódolt szakaszokból, exonokból ill szekvenciák  $\alpha = 0,50 \pm 0,01$  exponenssel jellemezhetőek, azaz nem mutatnak hossz tv korrelciákat. (Ezek a szekvenciák egyszerűen olyan, alacsony rendű fajokból származó genetikai szekvenciák, amelyek nem tartalmaznak intronokat; másrészt cDNS-szekvenciák, azaz (ld. pl. Weaver & Hedrick-nál) olyan dezoxi-ribonukleinsav-szekvenciák, amelyeket mRNS-ekből, reverz transzkriptáccal állítottak elő, abban a fázisban, amikor az elsődleges transkriptból (*primary transcript*-ből) az intronokat már kivették a megfelelő sejtszintű folyamatok.) Viszont az intronok is tartalmaznak szekvenciákat, azaz  $F(l)$  függvény exponense  $\alpha = 0,61 \pm 0,03$ . (Peng et al. [1992]) A kódot erősen szignifikáns eltérése még nincs kellő alapossggal megmagyarázva.

Írott szvegek esetében, a vizsgálathoz használt szvegek között megtaláljuk a Biblia eredeti, héber nyelvű szövegét, valamint modern fordításait, továbbá Shakespeare-dramákat, regényeket, sőt egy szöveget is. Néhány érdekesebb eredmény:

- A szvegek sok nagyságrenddel kisebb  $\alpha$ -exponenssel jellemezhetőek, melynek értéke általában  $0,6 - 0,7$  között esik (Schenkel et al. [1993]).
- A kódot nem jellemző a szerzőre, hiszen például a *Hamlet* exponense  $0,56$ , míg a *Romeo és Júlia*  $0,6$  (Schenkel et al. [1993]).
- A fordítások soraiban, gyökér, a korrelciák értéke csökken. Habár a Biblia exponense kimagaslóan magas ( $\sim 0,75$ ), a fordításai soraiban ezen értékek szisztematikusan csökkennek.<sup>6</sup> (Amit et al. [1994])
- A szöveg a szólag független szövegek hosszánál jóval hosszabb korrelciákat mutat, amely jelenség nem magyarázható pusztán a tartalom korrelációjával. Ezek szerint a szöveg szövegeinek elrendezése "nem véletlenszerű", előzetes sejtésünkkel ellentétben, írja (Schenkel et al. [1993]).
- A szöveget kisebb részekre, például szavakra, mondatokra vagy  $l = 10 - 10000$  karakter hosszúságú, azonos darabokra szabdalva s ezeket a darabokat szabadon permutálva, a szabdalás hosszának nagyságrendjéig megmaradnak a korrelciák, azon túl pedig eltűnnek (Schenkel et al. [1993]).

Lefordított számítógépprogramokat (.exe-file-okat) is megvizsgáltuk, ezek esetében  $0,9$ -et is meghaladó kódot találtak (Schenkel et al. [1993]). érdekes még a "humán véletlenszámgenerátor"

---

<sup>6</sup> Nem találtam a szakirodalomban világszerte arra a kérdésre, hogy a magas hatványkitevő miért nem lehet a héber nyelvnek vagy ortográfának a jellegzetes tulajdonsága.

vizsgálata: az egyik szerző 0 s 9 kztt írt le szmokat, "vletlenszerűen". A vletlenszerű eloszlásra val trekvs eredmnyekppen, rvid tvon ( $l \sim 10$ ) anti-korrelcit fedezhetnk fel, de ennl hosszabb tvon – a szmokat generl szemly minden igyekezete ellenre – megjelentek a korrelcik. Radsul, a hatvnykitevő rtke nem is lland, tart az 1-hez! A szerzők ezt gy magyarzzk, hogy – ellenttben a szvegek s a szmítgpes programok rsval –, a vletlenszmok generlsnak nincsen rtelmes clja, vagyis a tudattalan tnyezők, amelyeket felelőss tesznek a hossz tv korrelcikrt, nagyobb szerepet jtszhatnak (Schenkel et al. [1993]).

Az  $F(l)$ -fggvny szoros kapcsolatban ll a szimblumszekvencia, mint jelsorozat Fourier-transzformltjval is. Hossz tv korrelcik esetr az  $S(f)$  teljes tmnyspektrum szintn hatvnyfggvny:

$$S(f) \sim f^{-\beta}. \quad (3.15)$$

A kt exponens kztti elmleti sszefggs (Schenkel et al. [1993]):

$$\alpha = \frac{\beta + 1}{2}. \quad (3.16)$$

A DNS-szekvencik diszkr Fourier-analízisből kapott eredmnyek jl visszaadjk ezt az elmleti sszefggst is.

Miknt használhatk a hossz tv korrelcik a kdol s a nem kdol szakaszok klnvlasztsra? Stanley et al. [1993a] nyomn tekintsnk egy  $m$  hosszsg "megfigyelsi dobozt" (az említett cikkben  $m = 800bp$ ), s mozgassuk a dobozt a DNS-szekvencia mentn. Minden lpsben kiszmljuk a dobozon belli szekvencinak a fentiekben bevezetett  $\alpha$  exponenst, s brzöljuk ezt az rtket a doboz pozciójnak a fggvnyben. Az idzett cikk 9. brja illusztrlja ezen j algoritmus ltal produklt fggvnyt az lesztő III. kromoszmjnak feldolgozsaor: ahol psztz doboz kdol szakaszba lp, leesik a fggvnyrtk (az exponens, a pozci fggvnyben) 0,5-re, ahol pedig kilp a kdol szakaszbl, ott megnő az exponens 0,6 fl.

### 3.4 A klcsns informci-fggvny

*Ivo Große* s *Hanspeter Herzel* a megfigyelhető korrelcikát a kdol szakaszokban lvő klnbző kodonok nem egyenletes eloszlásra vezeti vissza (Herzel et al. [1997a], Herzel & Große [1995, 1997b], Große et al. 1997]). Bevezetnek egy "klcsns informci-fggvnyt" (*mutual information*), amelynek segítségével jobb algoritmust javasolnak a kdol s a nem kdol szakaszok klnvlasztsra.

Ezen eljrs megrtshez tekintsnk t az informcielmleti entrpiafogalom lnyegt a szimblumszekvencik szempontjbl (Shannon & Weaver [1986]).

Az informácielmélet Shannon-féle megközelítésnek a célja az, hogy elvi határt adjon arra, hogy milyen sebességgel lehet továbbítani valamely "nyelv" "mondatait" egy (zajment vagy zajos, véges sok állapottal rendelkező vagy folytonos) csatorná keresztül. Ebben a megközelítésben egy "mondat" hírt hordoz a valószínűségektől függően: minél valószínűtlenebb a mondat, annál nagyobb a hírt hordoz. (Gondoljunk csak az újságírók munkájára: az nem hír, ha a kutya megharapta a postást, de az igen, ha a postás megharapta a kutyát.) A biztos eseménynek nincs hírt hordoz, mert azt nem is kell továbbítani, elegendő, ha a vevő oldalon automatikusan hozzátesszük a vett zenéhez. A negatív logaritmusmal definiált entropia pedig az motiválta, hogy a két kapcsolattal, két lyukkréppel vagy két floppyval továbbíthat információ mennyisége kiegészítése legyen az egy esetben továbbítható.

Shannon, következő lépésben bevezeti a forrás  $H$  entropiáját:

$$H = - \sum_i p_i \log(p_i), \quad (3.17)$$

ahol  $p_i$  az  $i$ -ik esemény bekövetkeztének, az  $i$ -ik karakter előfordulásának a valószínűsége. Az entropia annál nagyobb, minél több esemény, minél homogénebb valószínűségekké válik el. Ez az oka annak, miért olvashatunk ritkán az újságban kutyaharapsort, de annál gyakrabban a lottház eredményéről: az előbbi esetben két, egy valószínűtlen és egy nagyon valószínű eseményről van szó, míg az utóbbiban sok, egyenlően valószínű eseményről.

Amennyiben a forrás nem egyszerűen független eseményeket állít elő, hanem egy korrelált szimbólumsorozatot (mondatokat) produkál, akkor a forrás entropiája:

$$H = - \lim_{n \rightarrow \infty} \frac{\sum_{Q \in \Sigma^n} P(Q) \log(P(Q))}{n}, \quad (3.18)$$

ahol  $P(Q)$  az  $n$  betűből álló  $Q$  sztring előfordulási gyakorisága. ( $\Sigma^n$  jelenti a  $\Sigma$  elemeiből alkotott,  $n$  hosszúságú sztringek halmazát.)

Az informácielmélet további elemeire nem lesz szükségünk. Az "entropia" elnevezés a fizikából származik. Az analóg kapcsolat miatt, további kapcsolat is kimutatható: egy fizikai entropia (definíciós konstans szorzóval eltekintve) megegyezik a részecske-trajektória sebességkoordinátájának informácielméleti entropiájával.

Ezen kitérő tesztet írta fel Herzog és Große [1995] *klcsns informáci-függvény* (*mutual information*). Jele a  $\lambda$  elemből álló  $\{A_1, A_2, \dots, A_\lambda\}$  betűk esetén  $p_{ij}(k)$  annak együttes valószínűsége, hogy valamely pozícióban az  $A_i$  szimbólumot találjuk,  $k$  pozícióval később pedig az  $A_j$  betűt. Jele  $p_i$  annak a valószínűsége, hogy egy  $n$ -készes  $n$  pozícióban az  $A_i$  szimbólum fordul elő, míg  $q_j$  annak a valószínűsége, hogy az  $n + k$ -ik helyen  $A_j$  található. Mivel a szekvenciákról feltesszük, hogy stacionárius,  $p_i = q_i$  minden  $i = 1, 2, \dots, \lambda$ -ra. Két szimbólumot akkor és csak akkor nevezünk *statisztikailag függetlennek*, mint arról már volt szó, ha  $p_{ij}(k) = p_i \cdot p_j$ . Az

$$I(k) := \sum_{i,j=1}^{\lambda} p_{ij}(k) \cdot \log \frac{p_{ij}(k)}{p_i \cdot p_j} \quad (3.19)$$

mennyiség, amelyet *klcsns informcinak neveznk*, a statisztikai függetlenséget, ill. a korrelációk erősségét méri.

Herzel és Große kutatásai szerint különböző lőnyek kódozott DNS-szekvenciákban megfigyelhető egy nagyon jellegzetes 3-hosszúságú oszcilláció a lassan lecsengő  $I(k)$  függvényben, még nagy távolságokon ( $k > 1000bp$ ) is: ha  $k$  hárommal osztható,  $I(k)$  többszöröse a szomszédos  $k-k$  esetén felvett  $I(k)$  értékeknek. Ennek oka a szerzők abban lálik meg, hogy eltérő az egyes bázisok gyakorisága a kódon (triplett) első, második, ill. harmadik pozícióban. Nem kódozott szekvenciák esetén ez a periódicitás nem figyelhető meg.

Größe et al. [1997] bevezetik az *AMI tagos klcsns informci* (*Average Mutual Information*) nevű mennyiséget, amely az  $I(k)$  függvény értékeinek tagja egy  $N$  hosszúságú (pl. a  $k \in [5; 5 + N]$ ) intervallumban ( $N \approx 50$ -től működik az eljárás). Nem kódozott szekvenciákra az AMI lecseng  $N^{-2}$ -vel, míg a kódozott szekvenciákra az AMI egy pozitív értékhez tart, amely érték a bázis-gyakoriságok kódon-pozícióval való függéséből származtatható.

Ezen megfigyelés egy olyan algoritmushoz vezet, amely a kódozott és a nem kódozott szekvenciákat a többi algoritmushoz hasonló megbízhatósággal képes szétválasztani, nem igényel előzetes tanítást (több algoritmus is létezik, amelyek a neuronhálózatoknál megszokott előzetes "training"-ből származó paramétereket használnak fel), és elég jó "finomsággal" ( $\approx 50bp$  hosszúságú doboz pontossággal) működik.

### 3.5 Redundancia

Az emberi nyelvek közismert tulajdonsága a redundancia. Ez teszi lehetővé, hogy a kommunikációt ne zavarják olyan "zajok", mint a beszélő beszédhibja, nyelvbotlás, rossz szöveg esetén az elgépelt, és így tovább. Ez a redundancia jellemző a genetikai kódra is, mint ahogy azt már megemlítettük.

Shanon *relatív entrpijának* nevezi egy forrás (3.17)-ben definiált entrpijának, és az azonos szimbólumok segítségével elérhető maximális entrpijának az arányát (Shanon & Weaver [1986]). Ez a mennyiség a maximális lehetséges kompresszió jelenti, ha ugyanazon bemenetből kódozunk. A *redundancia* a relatív entrpiát egészíti ki 1-re. Mantegna [1995b] jellesei szerint az *n-gramok entrpija* ( $\lambda$  elemű bemenet)

$$H(n) := - \sum_{i=1}^{\lambda^n} p_i \log_2 p_i, \quad (3.20)$$

ahol a  $p_i$ -k az egyes bzis  $n$ -esek valsznúságt jellik, mg a redundancia:

$$R := 1 - \lim_{n \rightarrow \infty} \frac{H(n)}{kn}, \quad (3.21)$$

ahol  $k := \log_2 \lambda$ . Mantegna megfogalmazsa szerint a redundancia a nyelv *flexibilitásnak* a megnyilatkozása (Mantegna et al [1994, 1995b]).

Nagy Ferenc [1986] becslésként 68%-ot számol egy tágos nyelv redundanciájára (32 betű, 10 betűs szavak esetén). Shannon a htkznapi angol nyelv redundanciáját durván 50%-nak becsli. Ezt az eredményt adja mind a nyelv Markov-lncokkal trtnő approximációjn alapul entrpi-aszmts, mind a titkosrsok elméletnek nhny eredménye. Ezt tmasztja al az is, hogy a ksrletek szerint egy angol szveg ltalban rekonstrulhat, ha a karakterek felt vletlenszerűen trljk. Emltsre mlts mg Shannon eszmefuttatsa arról, hogy milyen mrtkű redundancia esetén lehetséges keresztretjvnyt ksztetni: ha tl sok megszorsts vonatkozik a nyelvre, nyilván nehéz nagy keresztretjvnyt szerkesztetni. Viszont zrus redundancia esetén brmely betűsorozat ltező szt ad, vagyis a karakterek tetszőleges kt dimenzis elrendezése keresztretjvny. Shannon szerint 50%-os redundanciáig lehetséges kt dimenzis, mg 33%-os redundanciáig hrom dimenzis retjvnyt ksztetni, ha a nyelvre vonatkoz megszorstsok vletlen természetűek.

Mantegna et al. [1994, 1995b] klnfle szekvencikra, valamint ezen szekvencik kdol, ill. nem kdol szakaszainak konkatenciára szmtotta ki a (3.20) szerinti  $H(n)$  entrpit. Mivel rtkelhető eredményt csak  $L > 100 \times 4^n$  esetén kaphattak,  $n \geq 8$  rtkeket mr nem vizsgálhattak, sőt egyes szekvencik esetén mg kisebb  $n$  rtkeknl meg kellett llniuk. A (3.21)-beli konvergencia nagyon lass, s ilyen kis  $n$ -ek esetén a hosszabb repeat-szekvencik hatst sem vizsgálhatjuk. Ennek ellenre emltsre mlts eredményekre jutottak: nhny kivételtől eltekintve,<sup>7</sup> a nem kdol szekvencik redundanciája szignifiknsan magasabb a kdol szekvencik redundanciájnl, mg a komplett genomdarab grbje e kettő kztt halad az egyikhez kzelebb attl fggően, hogy a kdol vagy a nem kdol elemek vannak-e tbsgben az egysz genomban.

Mantegna et al. e megfigyelésből is azt a kvetkeztetst vonjk le, hogy a nem kdol szekvencik, szemben a kdol elemekkel, a statisztikai tulajdonságaik tekintetben a természetes nyelveken rott szvegekre hasonltnak.<sup>8</sup>

### 3.6 Modellezs generatív nyelvtanok segtsgvel

<sup>7</sup> A kivételként kzlt brbl kiderl, hogy ebben az esetben csak  $n \leq 4$ -re lehetett a  $H(n)$ -eket szmtani, s a grbk "trendje" alapján asszimpttikusan itt is lehetséges, hogy a fenti megfigyelés megllja a helyt.

<sup>8</sup> "Linguistic features of noncoding DNA sequences", mint azt az egyik cikkk cmben is emltek.

Az eddigiekben a DNS-szekvencik sok olyan statisztikai tulajdonsággal találkozunk, amelyek a természetes (emberi), valamint a számítógépes nyelvekkel rokonítják a DNS-szekvencikat. Arra is láttunk több rivet, hogy a szimblumszekvencik hagyományos sztochasztikus modelljei, a Markov-folyamatok nem rják le kielgőően az említett megfigyeléseket. Ezért az alábbiakban egy másik modellt fogok javasolni, a generatív grammatikák sztochasztikus általánosításait, valamint a Zipf-függvény elméleti szintjét is megmutatom, a modell paramtereiből. A generatív grammatikák elmélete s az automataelmélet nem sztochasztikus fejezetei egyenként az algebra bevett grnak szintjának, gy itt csupán a szükséges fogalmakra fogom korlátozni az ismertetést, amelyet az indokol, hogy e téma nem szerepel a fizikus szak tananyagban.

### 3.6.1 A formális nyelvek elméletnek alapfogalmai

A *formális nyelvek elméletének* a célja a természetes és számítógépes nyelvek leírása. Az alapgondolat az, hogy a nyelvet, mint a *”jólformált mondatok”* – azaz a nyelvhez tartozó, ábécébeli sztringek – halmazát, nem csak a halmaz elemeinek a felsorolásával, vagy a halmazok megadásánál megszokott, egyszerű szabályok segítségével definiálhatjuk. Megadhatunk egy nyelvet a *”szerkezetének”* leírásával is, *generatív grammatika* segítségével, valamint olyan automata révén, mely a nyelv *”mondatait”*, *”formuláit”*, és csak azokat fogadja el. Mindkét fogalom egy *véges vagy végtelen* halmazt ad meg, *véges* eszközökkel. Ezen gondolat mögött az áll, hogy a gyermek az anyanyelvét nyilván nem a hallott mondatok egyszerű reprodukálásával sajátítja el: egyrészt, nem hallhatja a nyelv összes, potenciálisan végtelen számú mondatát, és képes olyan mondatot is reprodukálni, amelyet előzőleg nem hallott; továbbá, egyszerű ismétlés esetén, nem hibázna, hiszen feltehetőleg csak helyes mondatot hall. Tehát – legalábbis Chomsky és követői szerint – a gyermek a nyelv szabályait sajátítja el.

A formális nyelvek elméletében, valamilyen  $L$  nyelv egy nemüres, véges  $\Sigma$  halmaz, az ún. *ábécé* fölött, definíció szerint nem más, mint a  $\Sigma$  elemeiből képzett, véges hosszúságú sztringek egy halmaza. (Az  $n$  hosszúságú sztring egy rendezett  $n$ -est jelent matematikailag.) Jelölje  $\Sigma^+$  a  $\Sigma$  elemeiből képezett, pozitív (véges) hosszúságú sztringek halmazát,  $\Sigma^*$  pedig ennek kibővítését az  $\epsilon$  (”empty”) üres, azaz nulla hosszúságú sztringgel. Ekkor egy  $L$  nyelv  $\Sigma$  fölött nem más, mint  $\Sigma^*$  egy részhalmaza:

$$L \subseteq \Sigma^*$$

Az alábbiakban *”szó”*, *”mondat”*, *”jelsorozat”*, *”formula”* kifejezéseket a *”sztring”* szinonímájaként fogom használni, a *”jel”*, *”betű”*, *”szimbólum”* szavak pedig az ábécé

elemekre fognak utalni. A szóhasználat a konkrét implementációtól függ: szintaxis (mondattan) esetén például  $\Sigma$  szavakat tartalmaz.

Két formula,  $X$  és  $Y$  *konkatenációján* azt a  $Z$  formulát értjük, amelyet úgy kapunk, hogy  $X$ -et és  $Y$ -t "egymás mellé írjuk", összefűzzük:  $Z = XY$ . Nyilvánvaló, hogy  $\Sigma^*$  a konkatenáció műveletével együtt egységelemes félcsoporthoz alkot, ahol az  $e$  üres szó játssza az egységelem szerepét.

A *generatív grammatika* fogalmának definíciója Révész [1979.] alapján az alábbi:

1. *Definíció:* Egy  $G$  generatív grammatikán az alábbi rendezett négyest értjük:  $G = (V_N, V_T, S, F)$ , ahol  $V_N$  és  $V_T$  diszjunkt véges ábécék,  $S \in V_N$ , az  $F$  pedig olyan rendezett  $(P, Q)$  pároknak egy véges halmaza, melyekre  $P, Q \in V^*$  ( $V := V_T \cup V_N$ ), és  $P$  legalább egy  $V_N$ -beli jelet tartalmaz.

A  $V_N$  elemeit *nemterminálisok* elemeknek,  $V_T$  elemeit pedig *terminális* jeleknek fogjuk nevezni, az alábbiakban ismertetendő okokból. Az  $F$  elemeit *helyettesítési szabályoknak* (*rewriting rules*) hívjuk, és  $P \rightarrow Q$  alakban írjuk. Az  $S$  kitüntetett nemterminális elem neve: *mondatszimbólum*. Az alábbi módon definiáljuk a *levezetés* fogalmát:

2. *Definíció:* Adott  $G = (V_N, V_T, S, F)$  grammatika esetén,  $X, Y \in V^*$ -ra azt mondjuk, hogy  $Y$  *levezethető*  $X$ -ből, azaz  $X \Rightarrow Y$ , ha létezik olyan  $P_1, P_2, P, Q \in V^*$ , hogy  $X = P_1 P P_2$ ,  $Y = P_1 Q P_2$ , valamint  $(P \rightarrow Q) \in F$ .

Ez azt jelenti, hogy ha  $X$ -ben  $P$ -t újraírjuk  $Q$ -val az egyik  $F$ -beli *helyettesítési szabály* alkalmazásával, úgy  $Y$ -t kapjuk. Jelöljük a  $\Rightarrow$  reláció tranzitív lezártját  $\Rightarrow^*$ -val, azaz  $X \Rightarrow^* Y$  csak akkor áll fenn, ha  $X$ -ből kiindulva,  $F$ -beli újraírószabályok nulla vagy véges számú alkalmazásával, eljuthatunk  $Y$ -ba. Egy  $G$  grammatika által generált  $L(G)$  nyelv az  $S$  mondatszimbólumból levezethető, terminális elemekből álló mondatok halmazát jelenti:

$$L(G) := \{P \in V_T^* \mid S \Rightarrow^* P\}$$

Azt mondjuk, hogy az  $L$  nyelvet a  $G$  grammatika *generálja*, ha  $L = L(G)$ . A terminálisok  $V_T$  halmazának szerepét a  $\Sigma$  ábécé tölti be, amikor egy nyelvhez grammatikát gyártunk. Két grammatikát *generálás szempontjából ekvivalensnek* nevezünk, ha ugyan azt a nyelvet generálják.

A formális nyelvek elméletének alapvető kérdése az, hogy milyen típusú nyelvekhez milyen grammatikákat adhatunk meg. Azaz milyen következményekkel jár, ha megkötéseket teszünk a helyettesítési szabályok alakjára. *Noam Chomsky* az alábbi *nyelvosztályokat* vezette be (Révész [1979.]):

3. *Definíció:* A  $G = (V_N, V_T, S, F)$  grammatikát  $i$ -típusúnak nevezzük, ha az alábbiak közül az  $i$ -ik teljesül:

$i = 0$ : Nincs semmilyen kikötés.

$i = 1$ : Az  $F$  minden eleme  $Q_1XQ_2 \rightarrow Q_1PQ_2$  alakú, ahol  $Q_1, Q_2, P \in V^*$ ,  $X \in V_N$  és  $P \neq e$ , kivéve esetleg az  $S \rightarrow e$  szabályt, amely viszont csak úgy szerepelhet az  $F$ -ben, ha az  $S$  nem fordul elő semelyik szabálynak sem a jobb oldalán.

$i = 2$ : Az  $F$  minden eleme  $X \rightarrow P$  alakú, ahol  $X \in V_N$ , és  $P \in V^*$ .

$i = 3$ : Az  $F$  minden eleme  $X \rightarrow PY$  vagy  $X \rightarrow P$  alakú, ahol  $X, Y \in V_N$ , és  $P \in V_T^*$ .

Valamely  $L$  nyelvet  $i$ -típusúnak mondunk, ha létezik hozzá  $i$ -típusú grammatika. Az  $i$ -típusú grammatikák osztályát  $\mathcal{G}_i$ -vel, míg az  $i$ -típusú nyelvek családját  $\mathcal{L}_i$ -vel szokás jelölni. Belátható, hogy ezen nyelvosztályok egymás valódi részhalmazai:

$$\mathcal{L}_3 \subset \mathcal{L}_2 \subset \mathcal{L}_1 \subset \mathcal{L}_0$$

Az  $i = 0$  nyelvosztályt *mondatszerkezetű* nyelveknek nevezzük. Az  $i = 1$  típusú grammatikák helyettesítési szabályai úgy néznek ki, hogy az  $X$  nemterminálist írjuk újra, a  $Q_1\_Q_2$  környezetben, ahol  $\_$  jelzi  $X$  helyét. Ezért az  $i = 1$  grammatika-, ill. nyelvosztályt *környezetfüggőnek* (*context-sensitive*, CS) nevezzük, szemben az  $i = 2$  *környezetfüggetlen* (*context-free*, CF) osztállyal. Az  $i = 3$  esetben pedig *reguláris* osztályokról beszélünk.

A formális nyelvekkel szoros kapcsolatban állnak az *automata-elméletből* ismert gépek. Egy nyelvet elfogad valamely automata, ha a nyelv mondatait, és csak azokat fogadja el. Az egyes nyelvosztályok ilyen alapon megfeleltethetőek az egyes automata-típusoknak. Erre a kérdésre – hely hiányában – nem fogok részletesen kitérni. Egyedül a véges állapotú automaták alap gondolatát ismertetem, mivel szoros kapcsolatban llnak a Markov-láncokkal.

Belátható, hogy reguláris nyelvekhez, és csak azokhoz szerkeszthető elfogadó *véges állapotú automata* (Révész [1979.]):

4. *Definíció:* Egy véges automatán az  $A = (K, T, M, q_0, H)$  rendezett ötöst értjük, ahol

$K$  egy véges, nem üres halmaz, az *állapothalmaz*,

$T$  egy véges ábécé, a *bemenő ábécé*,

$M$  a  $K \times T$  halmaznak egy leképezése a  $K$ -ra, az *átmenetfüggvény*,

$q_0 \in K$  a *kezdőállapot*,

$H \subseteq K$  a *végállapotok halmaza*.

A véges állapotú automata működése a következő: elindul a kezdőállapotból, az első lépésben beolvassa a *bemenő szalagra* írt mondat első jelét ( $\in T$ ), és a beolvasott jel,



valamint az éppen aktuális állapot függvényében, az  $M$  által meghatározott állapotba megy át. A következő lépésben újabb jelet olvas be a szalagról, és ez alapján újabb állapotba "ugrik" – feltéve, hogy az aktuális állapot és a beolvasott jel által alkotott rendezett pár eleme az  $M$  átmenetfüggvény értelmezési tartományának. És így tovább, egészen addig, amíg el nem akad az automata működése, vagy el nem olvasta az egész mondatot. Azt mondjuk, hogy az automata elfogadott egy mondatot, ha végigolvasta azt, és az utolsó állapot eleme  $H$ -nak. (Ez a folyamat formalizálható "állapotsorozatok" definiálásával, de ettől most tekintsünk el.)

A Markov-lncok lnyegben a vges lllapot automatk, illetve az azokkal ekvivalens "llapot-cimkzett automatk" (*state labeled automata*, Bird & Ellison [1994]) sztochasztikus ltalnos-tsai. Teht amikor eddig Markov-lncokkal kzeltettk a szimblumszekvencikat, regulris mod-elleket ksztettnk. Ezrt javasolom ebben a fejezetben azt, hogy regulris modellek (Markov-lncok) helyett a tgabb osztlyt jelentő, krnyezetfggetlen grammatikkat vizsgljuk.

A véges állapotú automata jellemzője, hogy nem rendelkezik "memóriával" (végtelen *veremmel*), és ebből következik majd a hosszútávú korrelációk hiánya a reguláris nyelveknél. Például az  $L_a := \{a^n b^n | n > 0\}$  nyelv (ahol  $a^n$   $n$  darab 'a' konkatenációját jelenti) azért nem lehet reguláris, mert a 'b'-k beolvasásakor "tudnunk kell" azt, hogy mennyi 'a'-t olvastunk be előzőleg. Márpedig, az 'a'-k száma tetszőlegesen nagy lehet, így ezt a végtelen sok lehetőséget nem tudjuk véges sok állapot segítségével "megjegyezni".

Ezzel szemben, a nem-reguláris környezetfüggetlen nyelvek tetszőlegesen sok "be-ágyazást" tesznek lehetővé. Tipikus példát szolgáltatnak erre – a fentebb megadott  $L_a$  halmazon kívül – a számítógépes nyelvek, ahol a ciklusszervezés jelenti az egymásba ágyazott struktúrákat, valamint a zárójelezett matematikai kifejezések. Chomsky [1957] amellet érvel, hogy az angol nyelv – és általában a természetes nyelvek – szintaxisa is környezetfüggetlen grammatikával adható meg, melyet a transzformációk átalakíthatnak. (Ellenpéldaként szokás felhozni egyes holland szerkezeteket, melyeket környezetfüggő szabályokkal érdemes csak leírni. Ezekről azonban tekintsünk el, mint ahogy azt a legtöbb nyelvész is teszi.)

A környezetfüggetlen nyelvek mondatainak jellemzője a látványos hierarchikus szerkezet. Egy formula levezetését egy gráffal (irányított fával) ábrázolhatjuk, melynek "gyökere" a mondat-simbólum, végpontjai ("levelei") a levezetett mondat szavai. A köztes csomópontok pedig a levezetés során megjelenő nemterminálisoknak felelnek meg, amelyből induló élek végpontjait "összeolvasva", azon helyettesítési szabály jobb oldalát kapjuk meg, amellyel újraírtuk a nemterminálist. Így például azt mondhatjuk (nagyon leegyszerűsítve a kérdés nyelvészeti oldalát), hogy az  $S$  mondat-simbólumból levezethető magyar mondat áll egy főnévi csoportból (NP), amely az alany szerepét tölti be, és egy igei csoportból (VP).

Azaz felírható a következő szabály:  $S \rightarrow NP VP$ . Maga az igei csoport is felépülhet igéből (V), tárgyból és határozókból (utóbbiak szintén NP-k, azaz főnévi csoportok):  $VP \rightarrow VP NP$ , illetve:  $VP \rightarrow V$ . A főnévi csoport pedig állhat főnevekből (N) és melléknevekből (A):  $NP \rightarrow A NP$ , és  $NP \rightarrow N$ . Majd a V, N és A nemterminálisokat helyettesíthetjük a megfelelő kategóriájú ("szófajú") magyar szavakkal (terminálisokkal). A fenti szabályok példát mutatnak a rekurziót lehetővé tevő szabályokra is, melyek segítségével lehet egy végtelen sok mondatból álló nyelvet leírni a generatív grammatikák véges eszközével.

A környezetfüggetlen nyelvekhez készített elfogadó automaták osztálya az ún. *verem-automaták*. Adott környezetfüggetlen grammatikához készíthető olyan automata is (ún. *parser*), amely a terminálisok lineáris szekvenciájából előállítja az azt létrehozó levezés(ek) során alkalmazott szabályok sorozatát.

Környezetfüggő nyelvre példa az  $L_b := \{a^n b^n c^n | n > 0\}$  nyelv. (A környezetfüggetlen nyelvekre szükséges feltételt kiróvó *Bar Hillel lemma* segítségével látható be, hogy  $L_b$ -hez nem adható meg környezetfüggetlen grammatika.) A környezetfüggetlenséget kizárja, ha a "korrelációk" egymást keresztezik, nincsennek egymásba ágyazva, mintha megengednénk a következő "zárójelezést":

$$(\dots[\dots]\dots)$$

Az  $\mathcal{L}_1$  nyelvosztály a *lineárisan korlátolt automaták* által elfogadott nyelvek osztályával egyezik meg, míg a mondatszerkezetű nyelveket elfogadó automaták éppen a híres *Turing-gépek*.

Környezetfüggő grammatikát használnak a fonológusok, a nyelvekben lejátszódó hangtani folyamatok jellemzésére. Viszont Kaplan és Kay [1994] bebizonyítja, hogy a fonológiai modellek általában olyan feltételeket rónak ki a szabályok alkalmazására, amelyek regulárisra redukálják a folyamatokat. Erre hivatkozva, az alábbiakban feltételezem, hogy a fonológia – elvileg – felírható reguláris szabályok segítségével is.

Láttuk, hogy a formális nyelvek elmélete éppannyira hasznos a nyelvészetben, mint a számítástechnikában. Úgy vélem, a genetikában is eredményesen lehetne felhasználni ezt a modellt. Hiszen a fehérjék hasonló hierarchikus szerkezettel rendelkeznek, mint a környezetfüggetlen nyelvek. Az elsődleges, másodlagos és harmadlagos szerkezet egymásra épülése, a funkciós csoportok elhelyezkedése, valószínűleg leírható ilyen eszközökkel, de saját magam – hiányos biokémiai ismereteim miatt – nem merek ezen a téren nyilatkozni. De úgy "érzem", hogy a fehérjeszerkezet-kutatások alapján felírhatóak olyan környezetfüggetlen újraíró szabályok, melyeket molekulafizikai számításokkal lehetne iga-

zolni. A gondolatmenetet folytatva, a következő kihívást azt jelentheti, hogy megértsük, ezek a környezetfüggetlen szabályok miként redukálódnak regulárisra, hiszen a c-DNS-ekből hiányzó korrelációk arra engednek következtetni, hogy a fehérjék "szintaxisa" nem csak környezetfüggetlen, hanem talán reguláris is egyben.

### 3.6.2 Sztochasztikus nyelvtanok

A generatív nyelvészet által használt formális nyelvek elméletének fentebb ismertetett formája nem teszi lehetővé, hogy a korrelációk statisztikus jelenségének nyelvészeti vonatkozásait megvizsgálhassuk, vagy a korrelációk létrejöttére nyelvészeti magyarázatot találjunk. Ehhez az szükséges, hogy a formális nyelvek elméletét kiegészítsük sztochasztikus eszközökkel. A környezetfüggetlen grammatikák ilyen általánosságait *sztochasztikus környezetfüggetlen grammatikák (SCFGs)* néven ismerik. A reguláris nyelvek osztálya sztochasztikus általánosságai között a Markov-modellek, több nyelvészeti általánossággal pedig nincs tudomásom. Utóbbiakra viszont egyelőre nincs is szükség, mivel a formális nyelvek elméletének legfontosabb felhasználói (nyelvészeti, számítástudományi) leggyakrabban az  $\mathcal{L}_2$  nyelvészettel dolgoznak.

Ha a generatív grammatikák négyesét egy valószínűségi függvénnyel egészítjük ki, sztochasztikus grammatikákat kapunk. Ilyen módon nem csak azt mondhatjuk meg, hogy egy adott mondat vajon eleme-e a grammatika által generált nyelvnek, hanem azt is, hogy milyen valószínűséggel vezethetjük le az  $S$  mondat-simbólumból az adott szekvenciát. Ezt a valószínűséget úgy értem, hogy amennyiben összefűzzük a "végtelen" sok mondatát, azaz képezzük egy *ideális szövegtörzset*, akkor az adott mondat milyen gyakorisággal fordul elő a törzsben. Az ideális törzs jó közelítést adhatja természetes nyelvek esetén egy könyv (pl. egy regény), a "DNS-nyelv" esetén pedig a DNS-szekvenciákat tartalmazó adatbázisokat fogom ilyen törzsnek tekinteni.

A nyelvészeti hagyományos szemléletmódja két ponton is eltér ettől a megközelítéstől. Először is, a nyelvi adatok helyes voltát binárisan kezeli: egy alak vagy grammatikus vagy agrammatikus, vagy eleme a nyelvnek, vagy nem. Ezen megközelítés hátrányait esetleg meglehetősen irónikus hangnemben, és a statisztikus-sztochasztikus szemlélet tükrözve írta le Abney [1996]. A másik eltérés pedig abban van, hogy a fonológiai grammatikát nem *körpuzsra (corpus-based)*, hanem a szókincsre, azaz a *lexikonra (lexicon-based)* vizsgálják, vagyis nem tesznek kísérletet azért, hogy például egy rendhagyó ("tiltott") hangkapcsolat egy gyakori vagy egy ritka szóban fordul-e elő. (Ennek az az oka, hogy a hagyományos generatív nyelvészeti mind a ritka, mind a gyakori szó a nyelv egyenrangú elemének tekinti, és az előfordulásban mutatkozó különbségeket a performanciához, azaz a nyelvészeti vizsgálati körre sorolja.)

A környezetfüggetlen grammatikák sztochasztikus általánosságait a következőképpen definiál-

hatjuk Krenn & Samuelsson alapján:

5. *Definíció: Sztochasztikus környezetfüggetlen grammatikának (Stochastic Context-free Grammar, SCFG) nevezzük az  $(V_N, V_T, S, R, P)$  ötöst, ahol:*

$V_N$  a nemterminálisok véges halmaza,

$V_T$  a terminálisok véges halmaza (legyen újból  $V := V_N \cup V_T$ ),

$S \in V_N$  a nemterminálisok közül a kitüntetett mondatzimbólum,

$R$  az  $X \rightarrow Q$  alakú levezetési szabályok egy véges halmaza, ahol  $X \in V_N$  és  $Q \in V^*$ ,

$P$  pedig egy  $R \rightarrow [0, 1]$  függvény oly módon, hogy

$$\forall X \in V_N : \sum_{Q \in V^*} P(X \rightarrow Q) = 1.$$

A  $P(X \rightarrow Q)$  valószínűség azt jelenti, hogy ha egy levezetési láncban megjelenik egy  $X$  nemterminális, azt milyen valószínűséggel fogjuk  $Q$ -ként újraírni.

Ezek után, egy levezetési lánc, ill. levezetési fa valószínűségét úgy definiálhatjuk, mint az alkalmazott helyettesítési (újraíró) szabályokhoz rendelt valószínűségek szorzatát. Valamely mondat valószínűsége pedig a hozzátartozó levezetési fák (*parse trees*) valószínűségeinek az összege lesz. A módszer a számítógépes nyelvészetben hasznos elemző automaták (*parser*-ek) futási idejének optimalizálására. Ha a korpuszt úgy tekintjük, mint nagy számú, egymás mellé írt mondatzimbólumból ("poli-S láncból") levezetett sztring, úgy látható a kapcsolat a korpuszból számított empirikus gyakoriság és a fentebb bevezetett elméleti valószínűség között.

### 3.6.3 Becslés a Zipf-függvényre

Ahhoz, hogy a modell paramtereiből eljuthassunk a Zipf-törvényben szereplő  $\omega(R)$  függvényhez, az alábbi gondolatmenetet javaslom: szmstjuk ki az egyes terminálisok előfordulási valószínűségét egy "korpuszban", azaz mondatok végtelen láncában. (Ezt a szmstst tudtommal még nem végezte el senki, mivel a sztochasztikus környezetfüggetlen grammatikát jelenleg még csak a parser-ek készítésére használjuk.)

$v_S(a)$ -val jelölöm azt, hogy az  $a \in V_T$  terminális várhatóan hányszor fordul elő egy  $S$ -ből levezethető mondatban:<sup>9</sup>

---

<sup>9</sup> Az alábbiakban a kisbetűk mindig terminálisra, a nagybetűk nemterminálisra, míg a görög betűk terminálisokból álló sztringre fognak utalni.

$$v_S(a) := \sum_{\sigma \in V_T^*} p_S(\sigma) \binom{\sigma}{a},$$

ahol  $p_S(\sigma)$  jelöli a  $\sigma \in V_T^*$  mondat fentebb bevezetett valószínűségét,  $\binom{\sigma}{a}$  pedig az  $a$  terminálisok számát a  $\sigma$  mondatban. (Megjegyzem, hogy habár formálisan nem látom be, de ezen végtelen szummának konvergensenek kell lennie. Hiszen felülről becsülhető a mondatok hosszának várható értékével, amely viszont véges, hiszen enélkül a kommunikáció nem lenne elképzelhető.)

A bennünket érdeklő  $v_S(a)$  várható értéknél többet fogunk kiszámítani a SCFG paramétereiből, azaz a  $P$  valószínűségfüggvényből. Jelölje  $p_A(\sigma)$  bármely  $A \in V_N$  nemterminálisra annak a valószínűségét, hogy  $A$ -ból a  $\sigma \in V_T^*$  sztringet vezetjük le: az  $A$  gyökerű,  $\sigma$ -t eredményező levezetési fák valószínűségeinek<sup>10</sup> az összegét. és jelölje  $v_A(a)$  annak a várható értékét, hogy az  $A$ -ból sztringekben hányszor szerepel az  $a$  terminális:

$$v_A(a) := \sum_{\sigma \in V_T^*} p_A(\sigma) \binom{\sigma}{a}. \quad (3.22)$$

Ezek a  $v_A(a)$ -k egy  $\mathbf{v}(a)$  vektornak a komponensei, amely egy, a  $V_N$  elemeinek a számával megegyező dimenziójú térben van.

A  $p_a(\sigma)$  átírásához be kell vezetni a *Chomsky-féle normálalak* fogalmát. Belátható (pl. ld. Révész [1979], SCFG-re Krenn & Samuelsson [1996]), hogy bármely környezetfüggetlen grammatikához létezik olyan Chomsky-féle normálalakban megadott grammatika (*Chomsky Normal Form, CNF*), amely ugyanazt a nyelvet generálja, és amelynek a szabályai

$$A \rightarrow BC$$

vagy

$$A \rightarrow a$$

alakúak, ahol  $A, B, C \in V_N$  és  $a \in V_T$ . Tetszőleges SCFG-hez is adható meg olyan SCFG, amelynek  $R$ -je CNF-alakú levezetési szabályokat tartalmaz, és amely minden sztringet ugyanolyan valószínűséggel generál, mint az eredeti SCFG.

Definiáljuk még a  $\pi(a)$  vektort és a  $\mu$  mátrixot:

---

<sup>10</sup> Az ilyen fák valószínűségét szintén az alkalmazott levezetési szabályok valószínűségeinek a szorzataként számíthatjuk ki, akár csak az  $S$  gyökerű fák esetén.

$$\pi_A(a) := P(A \rightarrow a)$$

$$\mu_{AB} := \sum_{C \in V_N} \left[ P(A \rightarrow BC) + P(A \rightarrow CB) \right]$$

Amikor az  $A$ -ból akarom levezetni a  $\sigma$ -t egy CNF alakú grammatikában, két lehetőségem van az elindulásra: amennyiben a  $\sigma$  egyetlen  $a$  terminálisból áll, úgy az  $A \rightarrow a$  szabályt kell alkalmaznom, egyébként pedig az  $A$ -t egy  $A \rightarrow BC$  szabállyal írom újra, és  $B$ -ből levezetem  $\sigma_1$ -et,  $C$ -ből  $\sigma_2$ -t, ahol  $\sigma = \sigma_1\sigma_2$  alakú (e kettő konkatenációja). Ennek megfelelően:

$$p_A(\sigma) = \sum_{b \in V_T} P(A \rightarrow a)\delta_{\sigma,b} + \sum_{B,C \in V_N} \sum_{\substack{\sigma_1, \sigma_2 \in V_T^* \\ \sigma = \sigma_1\sigma_2}} P(A \rightarrow BC)p_B(\sigma_1)p_C(\sigma_2) \quad (3.23)$$

(A  $\delta_{\sigma,b}$  szimbólum a megszokott Kronecker-deltát jelöli.) Ha beírjuk a (3.23) egyenletet (3.22)-be, úgy bizonyos egyszerűsítésekre lesz lehetőségünk, a következő három összefüggés felhasználásával:

$$\begin{aligned} \binom{b}{a} &= \delta_{b,a} \\ \sum_{\sigma \in V_T^*} p_C(\sigma) &= 1 \\ \binom{\sigma}{a} &= \binom{\sigma_1}{a} + \binom{\sigma_2}{a} \end{aligned}$$

Az eredmény a következő lesz:

$$v_A(a) = \pi_A(a) + \sum_{B,C \in V_N} P(A \rightarrow BC)(v_B(a) + v_C(a)),$$

amelyből:

$$v_A(a) = \pi_A(a) + \sum_{B \in V_N} v_B(a)\mu_{AB},$$

vagyis:

$$\mathbf{v}(a) = \pi(a) + \mu\mathbf{v}(a). \quad (3.24)$$

A (3.24) összefüggés átrendezésével a SCFG paramétereiből ki tudjuk számítani az egyes terminlisok frekvenciáit. Ehhez az szükséges, hogy az  $1 - \mu$  mátrix invertálható legyen, de ezzel azért nincs baj, mert a  $\mu$  elemeire nincsen semmiféle megkötés ( $\pi(\mathbf{a})$  kis megváltoztatására  $\mu$  is megváltozik), vagyis kis perturbációval megszüntethető egy esetleges szingularitás.

A hagyományos Zipf-trvny ppen a szintaxis, mint generatív grammatika, terminlisainak, azaz a szavaknak a most kiszámított gyakoriságait használja fel.<sup>11</sup> A karakter  $n$ -esekre végzett Zipf-analízisben szereplő frekvenciákra becsléseket kaphatunk, ha elég hosszú szavaink vannak, vagyis a szóhatron túli karakter- $n$ -esektől első közelítésben eltekinthetünk.

A következő fejezetben bemutatandó "vektortér-technika" – ahogy M. Damashek [1995], az alapötlet kidolgozója nevezte –, a karakter- $n$ -eseknek a Zipf-analízisben is szereplő, most megbecsült frekvenciáit használja fel.

---

<sup>11</sup> A hagyományos Chomskynus szintaxisban szerepelnek  $n$ . *transzformciók* is, amelyek egy generatív grammatika által generált nyelv szavait permutálják, de ezek a permutációk nem változtatják az általunk vizsgált gyakoriságokat.

## 4. fejezet

### A vektortr-technika

Az eddigiekben sok olyan statisztikai eszközzel megfogható jelenséggel találkozunk, amelyekben a kódol és a nem kódol (pontosabban: főleg ismeretlen szerepű, valószínűleg nem kódol) genom-részletek elternek. Feltűnő volt a hasonlóság a nem kódol DNS-szekvenciák és a természetes nyelven írott szövegek között. Foglalkozunk össze ezeket az eredményeket.

A nem kódol szekvenciák és az írott szövegek hatványfüggvényként lecsengő Zipf-függvényt eredményeztek. Ezzel szemben a kódol szakaszok Zipf-analízisekor exponenciálisakat kaptunk, amelyek az egyes betűk eloszlására emlékeztettek írott szövegekben. Hosszú tv korrelációkat, az írott szövegeken kívül, az intronokat tartalmazó génekben is találtunk. Az információméleti redundancia szintje e két típusú szimblumszekvenciában magas, szemben a "lehetőségeit" jól kihasználó exonokkal.

Az ilyen különbségekre például, exonokat kereső algoritmusokat a "mintázat-kereső" módszerek közé sorolnánk (ld. a 2.2 fejezetben), mivel egy előre definiált tulajdonságot (hosszú tv korrelációkat, hatványfüggvényt követő Zipf-plót, magas redundancia-részeket) keres. Az alábbiakban bemutatásra kerülő módszer ezzel szemben a "visszakereső algoritmusok" közé tartozik, amelyben a szekvenciák közötti hasonlóságot mérjük. Az eddig csak írott szövegekre alkalmazták, DNS-szekvenciákra még nem, így nem triviális az, hogy egy ilyen hasonlósági mérteket jelezni képes a funkcionális (kódol vs. nem kódol) különbségeket. Mégis, mint az ki fog derülni az eredményeimből, a javasolt módszer kiválóan különválasztotta az exonokat és az intronokat. A módszer sikere véleményem szerint összefügg Clay et al. [1996] és a többi, a 3. fejezet elején említett cikk eredményeivel, amelyek szerint hábr van korreláció az azonos eredetű exonok és intronok bázisfrekvenciái között, de azok szisztematikusan elternek egymástól.



## 4.1 A Damashek-ndszer ismertetse

Az elmúlt években egyre több algoritmus születik szövegek automatikus szelektálására: például egy hírügynökség számára nagyon hasznos lenne, ha a befutó híreket emberi beavatkozás nélkül lehetne nyelv és témakör szerint csoportosítani. Damashek [1995] éppen egy ilyen algoritmusra tesz javaslatot, melyet ő *Acquaintancenek* nevez.

Ennek az algoritmusnak a lényege az, hogy a szövegekből vektort készít, majd a vektorok skaláris szorzata jellemző a vektorok, azaz a szövegek hasonlóságára.

Képzeljük el, hogy egy  $n$  hosszúságú ablakot (a szekvencia s a szmtgpes kapacitves volta miatt ltalban  $n = 3..6$ ) mozgatunk végig a szövegen, karakterről karakterre. A különböző lehetséges karakter- $n$ -eseket indexeljük  $i = 1..J$ -vel. Jelöljük  $m_i$ -vel az  $i$ -ik karakter- $n$ -es előfordulásainak a számát.<sup>1</sup> A szövegből képezett  $\mathbf{x}$  vektor  $i$ -ik komponensét éppen az  $m_i$  lenormálásából kapott frekvenciaként képezzük:

$$x_i := \frac{m_i}{\sum_{j=1}^J m_j} \quad (4.1)$$

Ha van kt szövegnek, a belőlük ilyen módon képezhető vektorok,  $\mathbf{x}$  és  $\mathbf{y}$ , skaláris szorzata jellemzi a szövegek hasonlóságát:

$$S = \frac{\sum_{j=1}^J x_j y_j}{\left(\sum_{j=1}^J x_j^2 \sum_{j=1}^J y_j^2\right)^{1/2}} = \cos \theta \quad (4.2)$$

Fontos megemlteni, hogy a  $d(\mathbf{x}, \mathbf{y}) := 1 - S$  távolság nem definil metrikát a vektorok terében, azaz nem jó távolságfogalom a szövegek között. Ugyanis igaz, hogy  $d(\mathbf{x}, \mathbf{y})$  pozitív, szimmetrikus, s  $d(\mathbf{x}, \mathbf{y}) = 0$  akkor s csak akkor teljesl, ha  $\mathbf{x} = \mathbf{y}$ , mivel utbbiak az 1-es norma szerint lenormlt vektorok. (Attól a valószínűtlen esettől eltekintve, amikor két különböző szöveg ugyanarra a vektorra képezhető le, ez azt is jelenti, hogy a kt szveg azonos.) Ugyanakkor belthat, hogy a hromszg-egyenlőtlensg nem teljesl, pldul egy skban fekvő, kis szget bezr hrom vektorra.<sup>2</sup>

Ez a szorzat elsősorban a szövegek nyelv szerinti szétválasztására alkalmas, de szerencsés esetben az azonos nyelvű szövegek téma szerinti szétválasztása is lehetséges. Saját

---

<sup>1</sup> Fickett [1996] is emlt hasonl mrtket, de ő az  $n$  karakter hosszsg ablakot minden lpsben  $n$  karakterrel mozdítja el, vagyis a *nem* tfedő karakter- $n$ -esek frekvenciit használja.

<sup>2</sup> Legyenek pldul  $\mathbf{x}$ ,  $\mathbf{y}$  s  $\mathbf{z}$  egy skban fekvő vektorok ilyen sorrendben. A szomszdos vektorok zrjanak be  $30^\circ$ -os szget. Ekkor  $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{z}) = 1 - \frac{\sqrt{3}}{2} \approx 0,134$ , mg  $d(\mathbf{x}, \mathbf{z}) = 0,5$ , vagyis  $d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{z})$  nem teljesl.

*4.1. táblázat: Angol nyelvű fizikai (Ph), egyéb témájú angol szövegek (E), francia (Fr) és magyar (H) levelek (.txt-fájlok) alapján készített, a karakterhalmazok gyakorisággal jellemző ( $n = 3$ ) vektorok szorzata. Lásd, hogy az azonos nyelvű szövegek egymáshoz képest szignifikánsan magasabb szorzatokat adnak, mint a különböző nyelvű szövegek szorzatai. A módszer tényleg szerinti vizsgálatra is alkalmas, hiszen két E-szveg, ill. két Ph-szveg hasonlósága 15%-kal magasabb, mint egy E- és egy Ph-szveg szorzata.*

Magam például öt, fizikai témájú angol e-mailből (Ph1-Ph5), három szintén angol nyelvű, de más témájú e-mailből (E1-E3), két francia levélből (Fr1-Fr2) és három magyar nyelvű, magánjellegű e-mailből (H1-H3) álló korpuszt vizsgáltam. Az egyes szövegek hossza 3400 és 6000 karakter között volt, kivéve a két francia levelet, amelynek hossza csupán 1000 - 1200 karakter. Az angol ábécé 26 betűjét, a pontot, a vesszőt, valamint az üres helyet vettem figyelembe ( $r = 29$ -féle karakter). A többi karaktert ignoráltam, kis- és nagybetű között, ill. ékezetes és ékezet nélküli betű között pedig nem tettem különbséget. A több üres helyből álló szekvenciákat előzőleg törölni kell. Eredményeimet  $n = 3$ -ra és  $n = 4$ -re az 4.1, ill. a 4.2 táblázat tartalmazza.

A két táblázat adatai között szignifikánsan magasabbak az azonos nyelvű szövegek vektoraiból képezett vektorok szorzatai ( $n = 3$ -ra  $0,73 \pm 0,06$ ), mint a különböző nyelvűek szorzatai ( $0,25 \pm 0,024$ ). A két, különböző témájú angol szövegcsokor is elkülönül egymástól: a Ph-szövegek (pl.  $n = 3$ -ra a Ph-szövegek egymás közötti szorzata:

4.2. *tblzat: A 4.1 tblzatban hasznlt szvegek hasonlsga,  $n = 4$  karakterből ll "ablakot" hasznlva. A szorzatok rtke kisebb, az eltrsek mg szignifiknsabbak. Viszont nagyobb  $n$  csak hosszabb szvegekkel hasznlhat: a tbbi szvegnl jval rvidebb francia levelek hasonlsga mindkt tblzatban kisebb, mint a tbbi, azonos nyelvű szvegprok hasonlsga mrtke.*

$0,79 \pm 0,024$ ) ill. az E-szövegek szorzata ( $0,80 \pm 0,015$ ) magasabb, mint egy E- és egy Ph-szöveg szorzata ( $0,68 \pm 0,03$ ).

Damashek a módszert továbbfejleszti. Bevezeti az ún. *centroid vektort*, amely egy-egy szövegcsokor (például az angol nyelvű szövegek, vagy az angol nyelvű, számítástechnikai témájú szövegek) vektorainak az átlaga. Ezt a vektort levonva a szövegek vektoraiból, kitranszformálhatóak a szövegcsokor közös jellemzői, például az adott nyelv funkcionális-grammatikai szavaiból adódó jellegzetességek (a magyar esetén pl. az "□a□" hármas vagy az "□az□" négyes, ahol '□' az res karaktert, a *space*-t jelenti). Ilyen módon, a szövegek finomabb csoportosítása is lehetséges, tartalom, esetleg stlus szerint is.

A rvid szvegek statisztikai hibáinak a kikerülésére azt javasolja Damashek, hogy rendeljnk egy-egy nulla s egy kztti  $\lambda$  slyt az egyes vektorokhoz, kisebbet a rvidebbekhez s nagyobbat a hosszabbakhoz, majd a kt dokumentum ezen slyval, mint szorzfaktorial, korrigljuk a centroid vektorok levonst kvetően kapott szorzatot.

## 4.2 A Damashek-ndszer tovbbfejlesztse s diszkusszija rott szvegekre

A ndszert inkbb ms irnyba fejlesztettem tovbb: összefűztem mind a négy szövegtípusból (E, Ph, Fr, H) néhány szöveget egyetlen fűzérré.<sup>3</sup> Egy  $m = 100$  hosszúságú dobozt mozgattam végig ezen a fűzéren, és a doboz éppen aktuális tartalmából képezett vektort ( $n = 3$ ) összeszoroztam a fűzér elejéből (0-6775. karakterek közötti Ph-szövegekből) képezett vektorral. (Erre az eljárásra "mozg dobozos ndszer"-knt fogok hivatkozni a kvetkező fejezetekben.) Az eredményeket a 4.1. ábrán mutatom be. A fűzér angol és nem angol nyelvű részei élesen különválnak egymstl: ott, ahol a doboz nem angol szveget tartalmazott, drasztikusan leesett a doboz tartalmnak s az angol nyelvű referenciaszvegnek a hasonlsga. A ndszer alkalmasnak tűnik arra, hogy klnbző jellegű szvegekből sszetett szekvencikat a komponenseire bontsuk szt, legalbbis olyan pontossggal, amelyet a mozg doboz mrete lehetőv tesz.

Az eredményt rviden azzal magyarázhatjuk, hogy különböző nyelvekre különböző karakter-szekvenciák jellemzőek. Például az angol nyelvű szövegekben, az  $n = 3$  esetben kiugrik a "the" karakterhármas: gondoljunk a határozott névelőn kívül a névmásokra ("they", "their", "them", "these", "there", stb.).

Ezt a gondolatot kifejtve, a Damashek-ndszer sikert hrom tnyezőre vezethetjk vissza: kt nyelvi s egy nyelven kvli tnyezőre.

Az első tnyező szintaktikai-szemantikai: az adott nyelv mondattana meghatározza a nyelvre jellemző funkcionlis morfikat (nvelők, ktőszavak, előljk, nvmsok, ragok,...), mg a szveg tmja a gyakori, tartalommal br szavakat. Ezek a szvegre jellemző szavak erősen befolyoslják a dominns karakter- $n$ -eseket. Ahhoz, hogy a Damashek-ndszer sikernek ezen tnyezőjt rszletesen lerhassuk, szksg lenne a sztochasztikus generatív grammatikk nyelvsveti alkalmazsainak nagyobb mrvű elterjedsre, valamint a formlis szemantika eddig mg nem ltező sztochasztikus kiterjesztsre.

A msodik tnyező a nyelv fonológija, mghozz a hangtan *fonotaktiknak* nevezett ga, amely a markovi tulajdonságokat, vagyis a lehetsges hangkapcsolatokat rja le. A kutatsok ezen irnya előrbb tart, mint az előző bekezdésben említett krdsek, de mivel nem kvnok nyelvsveti problémkba belemerlni, csupn nhny gondolatot emltek.

A *fonotaktika* a fonológia azon ága, amely a fonémák *lehetséges* kapcsolatait vizsgálja a célnyelvben (Kiefer, [1994], 4. fejezet). Például a \**bárdás* szó nem létezik a magyar

---

<sup>3</sup> A file szerkezete: 0-6775. karakter: Ph-típus; 6776-13551. karakter: E-típus; 13552-23219. karakter: Ph-típus; 23220-25862. karakter: H-típus; 25863-32319. karakter: Ph-típus; 32320-35178. karakter: Fr-típus.

4.1. *bra*: Klnbző nyelvű s tartalm szvegeket fűztem, ssze, majd egy  $m = 100$  hosszsg dobozt mozgattam vgig ezen a fzren. Lpsről lpsre kiszmtottam a doboz tartalmnak s a sztring első 6775 karakterből ll, fizikai tmj angol szvegnek a hasonlsgt,  $n = 3$  mellett. Az *bra* a hasonlsg mrtkt (a Damashek-fle szorzatot) brzolja a doboz kezdőpozciinak a fggvnyben.

nyelvben, de nem érezzük idegennek, "akár létezhetne is", hiszen megfelel a magyar nyelv fonotaktikai szabályainak. Elvileg minden ltező sz, mint a magyar nyelv szkincsnek az eleme, jl formlt. Ez viszont rendkvl bonyolultt tenn a fonotaktikai szablyokat, hiszen sok hangkapcsolat csupn egyetlen szban fordul elő: pldul sz vgn [mt] nem fordulhat elő, kivve a *terem**t*** szban. Ezrt ezeket is agrammatikusnak tekintjk. Egy *corpus-based* statisztikus megkzelts klnbsget tud tenni a gyakori *terem**t***, *barack*, *recept* szavak kevsb agrammatikus

volta, a ritka s idegen hangzs **nganaszn**,<sup>4</sup> **scs**, **jl** fokozott agrammatikussga (amit az anyanyelvi beszlő is rez), valamint a nem előfordul hangkapcsolatok nyelvi sttusza kztt.

Tegyük fel, hogy a nyelvünk fonotaktikáját le tudjuk írni egy olyan Markov-moddellel,<sup>5</sup> amely  $N$  állapotot ( $s_1, s_2, \dots, s_N$ ) és  $M$  darab jelet ( $\sigma_1, \sigma_2, \dots, \sigma_M$ ) tartalmaz. Az  $i$ -ikből a  $j$ -ik állapotba történő átmenet valószínűségét jelöljük  $p_{ij}$ -vel, míg  $a_{ij}$  annak a valószínűsége, hogy az  $i$ -ik állapotban a modell a  $j$ -ik jelet bocsátja ki. Tegyük fel, hogy ergodik a Markov-moddellünk; és mivel hosszú láncokról, reguláris nyelv hosszú modatairól van szó (ne felejtsük el, hogy a reguláris nyelvek osztálya zárt a \* műveletre, ld. a 3.6 fejezetben), feltehetjük azt is, hogy a modell már "egyensúlyivá vált", azaz annak a  $v_i$  valószínűsége, hogy a szekvencia valamely pozíciójában a rendszer az  $i$ -ik állapotban lesz, független a pozíciótól. Ekkor a  $\sigma_{k_1}\sigma_{k_2}\dots\sigma_{k_n}$  szimbólum  $n$ -es elméletileg jóslt valószínűsége, azaz a Damashek-vektor megfelelő komponense:

$$P(\sigma_{k_1}\sigma_{k_2}\dots\sigma_{k_n}) = \sum_{i_1=1}^N v_{i_1} a_{i_1 k_1} \sum_{i_2=1}^N p_{i_1 i_2} a_{i_2 k_2} \dots \sum_{i_{n-1}=1}^N p_{i_{n-1} i_n} a_{i_n k_n}. \quad (4.3)$$

Mivel a  $p_{ij}$  és  $a_{ij}$  paramétereiket nem ismerjük, a jövőbeni kutatások feladata lehet valamely adott korpusz esetén ezen "rejtett Markov-modell" (*hidden Markov Model*, Krenn & Samuelsson [1996], Rabiner [1989]) paramétereinek a becslése. A becsléshez rendelkezésre állnak a  $P(\sigma_{k_1}\sigma_{k_2}\dots\sigma_{k_n})$ -k empirikus közelítései, valamint felhasználhatjuk a jelek empirikus gyakoriságait. Az  $i$ -ik jel  $\nu_i$  frekvenciájára adható elméleti becslés:

$$\nu_i = \sum_{j=1}^N v_j a_{ji}, \quad (4.4)$$

hiszen  $v_j$  valószínűséggel van a rendszer az  $s_j$  állapotban, és ekkor  $a_{ji}$  valószínűséggel bocsátja ki a  $\sigma_i$  jelet.

---

<sup>4</sup> Egy szamojd np s nyelv neve.

<sup>5</sup> Felmerlhet, hogy miért nem egyszerűsíttem le a gondolatmenetemet Markov-láncokra. Két okból ragaszkodtam a Markov-moddellhez. Egyrészt, a reguláris grammatikákat – amelyekkel lerhatjuk a fonolgit Kaplan & Kay [1994] szerint – Markov-moddellekre, és nem Markov-láncokra vezethetjük vissza. A második ok nyelvészeti: a fonotaktikai szabályok sok esetben felhasználnak metrikus fonológiai információkat is (szótagszerkezet, szószerkezet, stb.). Ezt úgy valósíthatjuk meg, ha a metrikus szerkezet elemeit (pl. szótagkezdet, mag, szótagzárlat) az állapotokkal reprezentáljuk, míg a szegmentumoknak (fonémáknak) felelnek meg a Markov-modell jelei.

Miután diszkutáltuk a Damashek-módszer sikerének két nyelvészeti tényezőjét, említsük meg a harmadik, nyelvészeten kívüli tényezőt is, a helyesírási hagyományt. Az "sz" pár csak azért jellemző a magyar szövegekre, mert a magyar helyesírás így jelöli a fonetikai [s] hangot. Ugyan az a [š] hang a magyar nyelvű szövegekben 's'-ként, az angolban leggyakrabban 'sh'-ként, a franciában 'ch'-ként, míg a németben 'sch'-ként jelenik meg. Eme tényező szerepe sem szabad figyelmen kívül hagyni, hiszen a német szövegekből készült vektorok meghatározó komponensei lesznek az 'sch'-t tartalmazó komponensek, amelyek ritkábbak a magyarban, szinte zűrszerűek lesznek. A helyesírási hagyomány hatását fonetikai-fonológiai trükkök alkalmazásával lehetne kiküszöbölni, de egyelőre nem látnak rendelkezésemre álló szövegekben elég hosszú ilyen típusú szövegeket.

### 4.3 A szövegszámítási algoritmus

A szövegszámítási algoritmusokat C-nyelven, UNIX operációs rendszer alatt programokkal végeztem el.

Először a feldolgozandó szövegszekvenciákat tartalmazó fájlokat készítettem el, e-mail leveleket mint .txt-fájlokat felhasználva, ill. a GenBank-ből letöltött genomrészletek megfelelő szövegeinek a megkereséssel. Az első program, amelyet *vektor* (vektor), ezekből a fájlokból – a paraméterként megadott  $n$ -re – elkészítette és fájlba mentette az  $n$ -gram-frekvenciákat tartalmazó vektorokat. Írott szövegek esetén a több, egymást követő részek karaktereit (*space*-eket) ki kellett hagyni, hiszen ezek a nyelvileg irreleváns ismétlődő szekvenciák indokolatlanul dominálhatnak a keletkező vektor jellegét, ill. a skalárszorzatokban kiugran magas, de a szövegek kategorizálása szempontjából nem indokolt tagokat eredményezhetnek.

A vektorok elkészítése során felhasználtam Damashek [1995] által javasolt trükköt. Mivel az  $n$  karakter hosszúságú ablakkal végrehaladtam a szekvencián, nem volt érdemes egy  $r^n$  méretű vektorba gyűjteni az egyes  $n$ -esek előfordulásait a szövegben (ahol  $r$  a szekvenciában előforduló karakterek száma, az "bc" mérete), hiszen főleg az  $L \approx 1000 - 5000$  hosszúságú szövegekre ( $r = 29$ )  $L \ll r^n$ . Damashek javasolta, hogy az egyes  $n$ -eseket találgattam egy-egy szöveg (egy  $r$  alap szövegszámítási rendszerben felírt szövegnek tekintve a karakter- $n$ -est), majd ezt az újabb szekvenciát rendeztem a "gyorsrendezés" (*quicksort*) nevű algoritmus segítségével (Press et al. [1989]).<sup>6</sup> Ezt követően már könnyű összehasonlítani és fájlba menteni a nullától eltérő gyakoriságokat (vektorelemeket).

Az így kapott fájlokat használta fel a *szorzás* nevű programom, amely a vektorok skalárszorzatát végezte el, és az eredményt a standard outputon jelentette meg. Írtam egy segédpro-

---

<sup>6</sup> Press et al. [1989] szerint tágosan a Quicksort a leggyorsabb rendezési algoritmus a legtöbb esetben, nagy számú elemre, és  $N \log N$  nagyságrendű a futási ideje. Ugyanakkor kihangsúlyozzák a szerzők, hogy a legrosszabb esetben igazából csak  $N^2$ -es az algoritmus.

gramot is (`planmk`), amely egy, a vektor-fjlokat tartalmaz listbl elkszt egy olyan vgrehajthat fjlt, amely a `szorz` program sokszori meghvsval mindegyik vektort szeszorozza az szes tbbivel.

A mozg dobozos technikt az `ablak` nevű program valstotta meg. Ez a program paramterknt kapja meg az  $n$  rtkt, valamint annak a fjlnak a nev, amellyel szoroznia kell a mozg doboz tartalmbl kpezett vektort. Ez a program a `vektor` s a `szorz` program lnyeges elemeit egyarnt tartalmazza, hiszen minden lpsben a doboz tartalmbl vektort kell kpeznie, majd azt szeszorozni a `szorz-fjllal`. A program olyan formtumban menti el az eredményeket, hogy az a `grapher` nevű adatfeldolgoz program szmra használhat legyen.

A korrelcimentes ("nav") vektorokat a `naivvekt` program ksztette el (ld. a 4.5 fejezetet), amely bemenő paramterknt felhasználta a *GenBank* fjljaiban szereplő bszgyakorisgokat. A mozg dobozos program egy fejlettebb vltozata (`window`) ugyanakkor, amely a mozg doboz ppen aktulis tartalmbl kpezett nav vektort s karakter- $n$ -es gyakorisgvektort szorozza ssze, minden egyes lpsben szeszmozolja a dobozban lvő szimblumokat fajtik szerint.

## 4.4 A felhasznlt adatbzis

Ahhoz, hogy Damashek-fle vektortrtechnikt DNS-szekvencikra alkalmazhassam, a <http://www.ncbi.nlm.nih.gov/> cmen tallhat *NCBI-GenBank* adatbankbl tltttem le a szksges adat-fjlokat. A *GenBankben* 1998. februrjban (*Release 105.0*) trolt adatok fajok szerinti megoszlst mutatja a 4.3. tblzat. Az *NCBI-GenBank* adatbzisa az elmlt vekben rohamosan fejlődtt (ld. a 2.1. tblzatot s a 2.1. brt), gy az adatbzis teljes feldolgozra, de mg egy jelentős rsznek a feldolgozra sem vllalkozhattam. Az albbiakban bemutatand eredmények teht egy meglehetősen szűk szekvencia-halmazra plnek, de arra utalnak, hogy rdemes a munkt folytatni, esetleg biolgiailag is motivlt krdek felttelvel.

A szekvencik vlasztsnl a fő szempont az volt, hogy az llatvilg klnbző evolcis szintjein ll fajok egyarnt kpviselve legyenek. Így felhasznltam az Epstein-Barr vrus genomjt (a *GenBankben* a `EBV.SEQ` nv alatt tallhat meg), az *E. coli* baktrium genomjnak egy rszlett (`ECOUW87.SEQ`), a *Caenorhabditis elegans* nevű hengeresfreg szekvenciit (`CELTWIMUSC.SEQ`), az lesztő mind a tizenhat kromozmj (egy rgebbi vltozatban `SCCHR.III.SEQ` a III. kromozmjt jelli, ill. az jabban az szes kromozma `yeast_chr_n.gb` nv alatt tallhat meg, ahol  $n = 01, 02, \dots, 16$ ), a *Ratus norvegicus* patknyfaj *gamma-crystallin* gnjt (`RATCRYG.SEQ`), valamint a *Homo sapiens* bta-globulin gnjt (`HUMHBB.SEQ`).

Az egyes szekvencikat tartalmaz fjlok tartalmazzk a szksges tudnivalkat is az adott szekvencival kapcsolatban: a szekvencia pontos meghatrozst, a faj pontos rendszertani besorolst, bibliogrfaiai adatokat, a ngy bzis arnyt, valamint az egyes rszletekkel kapcsolatos is-



ttelek száma	bzisok száma	fajnv
1111188	608930616	<i>Homo sapiens</i>
306659	145170295	<i>Mus musculus</i>
76171	121890190	<i>Caenorhabditis elegans</i>
61591	52924236	<i>Arabidopsis thaliana</i>
31719	33127199	<i>Drosophila melanogaster</i>
10487	28590166	<i>Saccharomyces cerevisiae</i>
4772	17437454	<i>Escherichia coli</i>
10847	14248318	<i>Rattus norvegicus</i>
25736	12834422	<i>Fugu rubripes</i>
27455	10878861	<i>Oryza sativa</i>
1067	9816926	<i>Bacillus subtilis</i>
21253	9381753	<i>Human immunodeficiency virus type 1</i>
1263	6706922	<i>Schizosaccharomyces pombe</i>
4630	5245512	<i>Gallus gallus</i>
602	4878602	<i>Mycobacterium tuberculosis</i>
11684	4416232	<i>Brugia malayi</i>
10825	4376759	<i>Toxoplasma gondii</i>
4790	4269092	<i>Bos taurus</i>
4135	3833814	<i>Plasmodium falciparum</i>
149	3830822	<i>Synechocystis sp.</i>

4.3. tblzat: A NCBI-GenBank tartalma 1998. február 15-n (Release 105.0), a DNS- s RNS-szekvencikat beleszmtva, de a mitokondrilis s a kloroplasztiszbl szrmaz szekvencikkal kihagyva.

mereteket (exonok, intronok, *primary transcript*-ek, repeat-szekvencik elhelyezkedése, mutcis pontok, s gy tovbb). Utbbi alapjn vlasztottam ki a felhasznlt rszszekvencikat. A clom a vlaszts sorn az volt, hogy minl hosszabb s minl kevsb vitatott szerepű rszleteket vlasszak ki. Azokat a szakaszokat is figyelmen kvl hagytam, amelyek a DNS-*kd* komplementumt tartalmazzk ("complement").

A vizsglt szekvencik kzl az alsbb rendű fajokhoz tartozk nagy mennyisgben tartalmazznak *kdol* szakaszokat (CDS-eket). Az lesztő kromoszmi kzl nhny nem tartalmaz intront, de tbbjkben tbb is tallhat, amelyek kzl a leghosszabbakat sszegyűjttem, akrcsak nyolc hossz CDS-t (ld. a 4.4. tblzatban). A *yex2* exont tartalmaz fjllhoz hozzadtk egy 135 bp hosszsg *repeat*-elem 14 pldnyt.

A kt emlős fajtl szrmaz gn meglehetősen komplex. Mindkettő 5-6 *primary transcript*-et, azaz a DNS-ről az mRNS-re trd szakaszt tartalmaz, kzte nagy szm, nem *kdol* szekvencival. Egy-egy ilyen *primary transcript* szerkezete is jellegzetes (ld. mg a 4.5 tblzaton): a vezrszakaszt, amely ltalban a transzkripci szablyozsrt felelős, kveti hrom exon, kzt k intronnal; utbbiak kzl a msodik jval nagyobb valamennyi említett szakasznl; vgezetl a

4.4. *tblzat: Azlesztő genomjnak felhasznlt rszletei. A nyolc darab exont yexn, mg a nyolc darab intront yinn jelli (n = 1.8).*

4.5. *tblzat: A HUMHBB.SEQ primary transcript-jeinek szerkeze- te: a vezrszakaszt kveti a hrom exon, kztik intronok, s vgl a vgszakasz zrja le a szekvencit.*

vgszakasz zrja le a *primary transcript*-et.

## 4.5 A vektortr-technika alkalmazsa DNS-szekvencikra

Az ismertetett eljrsok s szmtgpes programok segtsgvel elksztettem az egyes DNS-

4.6. táblázat: Az lesztő néhány exonjának s legnagyobb intronjainak hasonlósága,  $n = 4$  mellett. A nyolc exon exonból készített vektorok szorzata  $0,87 \pm 0,11$ , az exonok s az intronok hasonlósága  $0,76 \pm 0,095$ , míg az intronok egymás között csupán  $0,73 \pm 0,043$  hasonlóságot mutatnak.

szekvenciák Damashek-féle vektorait, különböző  $n$ -ek mellett, majd elvégeztem a megfelelő szorzásokat. Tekintsük tehát az érdekesebb eredményeket.

Az lesztő genomjából kiemelt, a 4.4. táblázatban felsorolt exonok s intronok szorzatait mutatja a 4.6. táblázat, amikor  $n = 4$ -eseket használtam. A táblázat szerint meglehetősen magas az exonok egymás közötti hasonlósági mértéke ( $0,87 \pm 0,11$ ). Ez az érték még magasabb lenne, ha eltekintünk a yex2 exonról, amely, mint rámutat, nagy mennyiségben tartalmaz ismétlődő szekvenciákat. Az exonok s az intronok hasonlósága mintegy egy szerszénnyival, közel 15%-kal kevesebb,  $0,76 \pm 0,095$ . A két adathalmaz egyértelműen elkülönülne, ha kihagynánk a yex2 exont. Ezzel szemben, az intronok egymás közötti hasonlósága nem jelentős, csupán  $0,73 \pm 0,043$ . Figyelemre méltó viszont a kis szerszén.

A HUMHBB s a RATCRYG-szekvenciák mindegyik *primary transcript*-jéből két "szövet" készítem: az egyik a három exon összefűzése, míg a másik a vezrszakasz, a két intron s a végszakasz, azaz a fehérje nem kódoló részek konkatencija. A 4.7. táblázat a HUMHBB-szekvenciának a 4.5. táblázatban felsorolt két *primary transcript*-jéből,  $n = 3$  mellett generált, kódoló (E1-E5) s nem kódoló (I1-I5) szakaszokat reprezentáló vektorainak a szorzattal mutatja. A vizsgált két szekvencia esetén kétféleképpen, hiszen az exonok egymással  $0,94 \pm 0,016$  hasonlóságot mutatnak, a nem kódoló szakaszok egymással való hasonlósága  $0,92 \pm 0,03$ , ugyanakkor a kódoló s a nem kódoló szakaszokból készített vektorok szorzata csupán  $0,75 \pm 0,08$ . Figyeljük meg a kis szerszén az első két esetben, ami azért is meglepő az exonok esetében, mert ezek rövid szekvenciák. Hasonló eredményeket kaphatunk más  $n$ -ekre, valamint a RATCRYG-szekvencia *primary transcript*-jeire is.

	E1	E2	E3	E4	E5	I1	I2	I3	I4	I5
E1	1,00	0,92	0,92	0,95	0,93	0,83	0,77	0,74	0,83	0,90
E2		1,00	0,97	0,93	0,95	0,73	0,66	0,62	0,73	0,85
E3			1,00	0,94	0,94	0,73	0,65	0,61	0,71	0,83
E4				1,00	0,95	0,78	0,71	0,67	0,79	0,87
E5					1,00	0,76	0,69	0,64	0,77	0,86
I1						1,00	0,92	0,90	0,94	0,93
I2							1,00	0,98	0,91	0,88
I3								1,00	0,90	0,86
I4									1,00	0,94
I5										1,00

4.7. *tblzat: A HUMHBB szekvencia t primary transcript-jból kpezett exon-fzrek (E1-E5), ill. nem kdol fzrek (I1-I5) hasonlsga, n = 3 mellett. A kt szekvencia-tpus lesen elvlik egymstl:  $E \times E = 0,94 \pm 0,016$ ,  $I \times I = 0,92 \pm 0,03$ , mg  $E \times I = 0,75 \pm 0,08$ .*

A kvetkező lpsben bevezettem a "valamely szekvencia nav vektora" vagy "korrelcimentes vektora" fogalmat: ennek  $i$ -ik komponensét gy kapom meg, hogy szeszorzom az  $i$ -ik bzis- $n$ -est alkot bzisoknak az adott szekvenciben megfigyelhető frekvenciit. Azaz, ha az  $i$ -ik  $n$ -gramm (bázis- $n$ -es) alakja  $\sigma_{k_1}\sigma_{k_2}\dots\sigma_{k_n}$ , és  $\nu_j$  jelöli a  $\sigma_j$  bázispár megfigyelt előfordulási frekvenciáját a vizsglt szekvenciben, akkor az  $\mathbf{x}^{(naiv)}$  vektor  $i$ -ik komponense:

$$x_i^{(naiv)} := P^{(naiv)}(\sigma_{k_1}\sigma_{k_2}\dots\sigma_{k_n}) := \prod_{j=1}^n \nu_{k_j}. \quad (4.5)$$

A 4.8 táblázatban lthat a kt emlős gn (HUMHBB s RATCRYG) t-t kdol, ill. nem kdol szakaszbl kpezett Damashek-vektornak, valamint a RATCRYG-szekvencibl kpzett korrelcimentes vektornak a szorzata ( $n = 4$ ). Kiolvashat, hogy a kódoló szakaszok jóval távolabb esnek a nav vektortól, mint a nem-kódolók: a szorzat értéke az előbbi esetben  $0,66 \pm 0,05$ , míg az utóbbiban  $0,83 \pm 0,03$ . Ez az eredmény nem mond ellent annak, hogy hosszútávú korrelációk ppen az intront tartalmazó szekvenciákban fordulnak elő, míg a tisztán kódoló szakaszokban nem, hiszen most a rvid tv korrelcik erőssgt vizsgltuk. A nem kdol szakaszokra kapott eredmny fajtl fggetlen, mg a patkny exonjai kicsivel magasabb szorzatot mutatnak, mint az emberi exonok. Utbbi eredmny nem magyazhat trivilis mdon azzal, hogy a *Ratus norvegicus* gnjből kpezték a korrelcimentes vektort, hiszen az exonok nagyon kis hnyadt jelentik az egész gn mretnek.

Ezekre az eredményekre alapozva, a 4.2. fejezetben ismertett "mozgó doboz"-os módszert alkalmaztam a DNS-szekvenciákra is, mghozz a HUMHBB *primary transcript*-jeire (4.2., 4.3. bra). A szorzat fjl az egész HUMHBB-szekvencia alapjn ksztt nav vektor

	kdol	nem kdol
RATCRYG 1	0,68	0,79
RATCRYG 2	0,71	0,87
RATCRYG 3	0,70	0,84
RATCRYG 4	0,70	0,86
RATCRYG 5	0,68	0,80
HUMHBB 1	0,69	0,84
HUMHBB 2	0,59	0,83
HUMHBB 3	0,59	0,80
HUMHBB 4	0,63	0,85
HUMHBB 5	0,62	0,81

4.8. *tblzat:* A RATCRYG s a HUMHBB gnek  $t$ - $t$  kdol, ill. nem kdol darabjnak hasonlsga a patkny-szekvencibl kpzett nav vektorhoz. A nem kdol rszek hasonlsga fajtl fggetlenl  $0,83 \pm 0,03$ . A kdol rszek esetn a szorzat 25%-kal, tbb szrssal alacsonyabb: a RATCRYG-exonok esetn kicsit magasabb, mint a humn exonok esetben.

4.2. *bra:* A G-gamma globulin *primary transcript*-jnek a krnyke az emberi HUMHBB-gnben, amely hrom exonbl (E) s kt intronbl (I) ll. Egy  $m = 100$  hosszsg dobozt mozgattam vgg a szekvencin, s lpsról lpsre brzoltam a doboz tartalmbl kpezett Damashek-fle vektornak az egysz HUMHBB-szekvencibl szmtott nav (korrelcimentes) vektorral kpezett szorzatt, a doboz pozicijnak a fggvnyben ( $n = 4$ ). Az exonok helyn leesik a szorzat, s megjelenik egy mly vlgy a *primary transcript* ktharmadnak a krnykn is.

volt. A HUMHBB átírásra kerülő részeinek jellegzetes struktúrájt mutatta be a 4.5 *tblzat*. Amikor a doboz rinti valamelyik exont, a szorzat rtke leesik, akrcsak a 4.1. *brn*, amikor

4.3. *bra:* A HUMHBB emberi  $\gamma$ -delta-globulint kódozó részét hasonlítottam össze a HUMHBB egységéből képzett korrelációs vektorral.  $n = 3$ ,  $m = 200$ . A pontozott pozíciókban tartalmaz a doboz exon-részletet: érdemes megfigyelni a két első völgy átlapolását.

a doboz nem angol nyelvű szövegbe lép. Így az exonokat egy völgy jelzi a hasonlóan mértékű a doboz kezdőpozíciójának a függvényben brzódiagramon: a doboznak minél nagyobb része tartalmaz exont, annál kisebb a szorzat értéke. Ha  $m$  a mozgó doboz szélessége, akkor a völgy  $\frac{m}{2}$ -vel balra van tolva az exonhoz képest, hiszen a völgy  $m$  balra az exon előtt elkezdődik, amikor a doboz egyik vége "belelép" az exonba, és az exon végén fejeződik be, amikor a doboz másik vége is elhagyja az exont.

A három exon közül az első kettő nagyon közel van egymáshoz, így van olyan helyzet, hogy a doboz belelép mindkettőbe. A *primary transcript* első két exonja körüli völgy tehát átlapol, viszont a harmadik exon völgye szépen kivehető. Meglepő felfedezés ugyanakkor a harmadik exont megelőző völgy feltűnése a *primary transcript* kódozó részében, mind az öt átírással kerülő szakasz esetén ugyan ott. Ezen "ál-exon" létrejöttére még nem találtam magyarázatot.

Végül, a 4.4 ábrán az élesztő III. kromoszómájának (SCCHRIII) egy részlete látható, az eddig bemutatott mozgó dobozos technikával. Ugyan ezt a részletet mutatja be Stanley et al. [1993b] 3. ábrája (a *DNA-walk*  $\alpha$ -exponense egy  $m = 800\text{bp}$  széles mozgó dobozban). Mindkét ábrában lokális minimumok jelzik az exonok helyét (ld. a 3.3. fejezet végén a részleteket), ennek ellenére a két ábra kevés hasonlóságot mutat.

4.4. *bra: Az lesztő III. kromoszmjnak a rszlete. A mozg doboz hossza  $m = 100$ , a szorz vektor az egysz kromoszma alapjn kpezett nav vektor. ( $n = 3$ )*

## 5. fejezet

### sszefoglals

Dolgozatom clja azon mdszerek bemutatasa volt, amelyek segtsgvel DNS-szekvencik statisztikai jellemzzeit vizsglni lehet. Mivel a genetikai kd ugyan olyan szimblumszekven- cia, mint a természetes nyelveken rdott szvegek (vagy akr a szmtgpprogramok), llandan felvetődik annak a lehetősge, hogy az egyik adathalmazra alkalmazott eljrst a msikra is kiprbljk. Így pldul a Zipf-analízist vagy az informcielmleti redundancia mrst a természetes nyelvekre alkalmaztk elsőknt, mg a hossz tv korrelcik kimutatsa "vletlen bolyongs" segts-

gvel előbb a genetikában történt meg.

A 3. fejezetben ismertetett módszerek alkalmazása a következő eredményekre vezetett:

1. A kódol és a nem kódol DNS-szakaszok több tekintetben is eltérő viselkedést mutatnak. Viszont sok hasonlóságot fedezhetünk fel utóbbiak és a természetes nyelveken rótt szövegek között.
2. A nem kódol szakaszok megfigyelt statisztikai tulajdonságai nem rható le Markov-folyamatok segítségével. Ezért javasoltam a sztochasztikus környezetfüggetlen grammatikák alkalmazását mindkét típus szimblumsorozat esetében, és ismertettem ennek a modellnek az alapfogalmait. Kiszámoltam a modell paramtereiből az egyes terminálisok előfordulási valószínűségeit, amelyre a Zipf-törvény kapcsolni lesz szükséges. (Ezt a számítást tudtommal még nem végezték el, mivel ezt a sztochasztikus módszert jelenleg még csak a kódokra alkalmazzák.)

Ezt követően, egy olyan eljárást alkalmaztam DNS-szekvenciákra, amelyet eddig csak rótt szövegekre használtak, ott viszont az algoritmus eredményesen választja szét a különféle szövegeket nyelvről és tartalom szerint. Az eredményeim azt mutatták, hogy a módszer az exonok és az intronok különválasztására is alkalmas:

1. Az evolúciós különböző szintjein áll fajoknál egyaránt megfigyelhető volt az, hogy az exonok egymással szignifikánsan magasabb hasonlóságot mutattak, mint az exonok és az intronok.
2. Két emlős fajnál szórakoztatónak is megmutattam, hogy az intronokban gyengébbek a rövid távú korrelációk, mint az exonokban.
3. A "mozgásdobozos" technikával megvizsgálva egy emberi gént, az exonok jellegzetes változatokat produkáltak. De ez a módszer jelenlegi állapotban még nem alkalmas az exonok megkeresésére egy adott szekvenciában, mivel még szakaszok is eredményezhetnek hasonló változatokat.

A módszer előnye az egyszerűség. Egyszerű egyszerű számításon, vagyis nem igényel se nem nagy számítási kapacitást, se nem előzetes "tréning"-et, szemben mégis, például neuronhálózatos alkalmazásokkal. Mésrészt maga az algoritmus is jól ttekinthető, és ezért könnyen magyarázható az, hogy milyen tényezőkre megy vissza a sikere. A módszer a szekvenciák hasonlóságát a karakter- $n$ -esek gyakoriságánál számítja:

- Írótt szövegek esetében az eltérő gyakoriságok oka lehet az eltérő téma (azaz még tartalmaz szavak gyakoriak a két szövegben) vagy az eltérő nyelvről. Utóbbi esetben a szintaxis, a fonológia és a helyesírás egyaránt felelős a karakter- $n$ -esek eltérő gyakoriságairól.

- DNS-szekvenciák esetében pedig az ok abban (is) keresendő, hogy Clay et al. [1996] szerint adott fajnál eltérő az intronok és az exonok GC-szintje, amely tényező már nem magában is elegendő lenne az intronok és az exonok kimutatott különbözőségnek a magyarázatára.



Zrsknt felmerl a krds, hogy ha a genetika s az emberi nyelvek kztt ilyen nagy fok hasonlsgokat fedeztnk fel, vajon beszélhetnk-e valamilyen rtelemben "genetikai nyelvről"?

Mantegna et al. [1995b] "nyelv" alatt procedurk s / vagy utastsok sorozatt rti, amelyek egy jl defnilt nyelvten szerint vannak elrendezve. Utal arra, hogy az ismert *genetikai kdon* tl, amely csupn a proteinek elsődleges szerkezetéről tartalmaz informcit, lehetsges, hogy lteznek egy "genetikai nyelv", amely a sejtbeli folyamatokra nzve tartalmaz utastsokat, s amelynek a genetikai kd csupn egy rszt alkotja. Egy kd statisztikai tulajdonsgai csupn a kdolt statisztikai tulajdonsgaitl fggenek, ezzel szemben egy nyelv statisztikai tulajdonsgait meghatározza a kommunikci egész rendszernek mgttés szerkezete. Ugyanakkor a cikk vge hangslyozza, hogy semmifle bizonytkot nem talltak a "genetikai nyelv" lte mellett.

Mantegna nyelvfilozfiai mlysgú nyelv-fogalomval szemben Ficket [1996] a gnek "szintaxisbl" hoz nhny pldt. Olyan szablyokat sorol fel, mint pldul mivel kell kezdődnie s vgződnie egy-egy exonnak, ha az a gn első, kzblső vagy utols exonja, vagy hogy hogyan pl fel egy *primary transcript*. Ezek a szablyok nyelvsveti rtelemben teszik nyelvv a DNS-szekvencikat.

gy sejtem, hogy a Ficket-fle "nyelvsveti genetika" s a nyelvsvet sztochasztikus modelljei jelentik azt a kt fő csapsirnyt, amelyen haladva eljuthatunk az említett nyitott krdekek megvlaszolshoz. Mantegna felvetsre pedig a krdekek tbbsgnek megvlaszolsa utn tudunk csupn felelni: ha ismerni fogjuk a genetikai szekvencik termszett, csak akkor mondhatjuk meg, milyen rtelemben beszélhetnk a "genetika nyelvről".

## Irodalomjegyzék

S. Abney [1996]: Statistical Methods and Linguistics

L. A. N. Amaral, S. V. Buldyrev, S. Havlin, M. A. Salinger, H. E. Stanley [1997]: Power law scaling in a system of interacting units with complex internal structure (preprint).

M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, N. Shnerb [1994]: Language and Codification Dependence of Long-Range Correlations in Texts, *Fractals*, **2**, 1, pp. 7-13.

Berend Mihly [1980]: Genetikai bc, Gondolat Kiad, Budapest.

S. Bird, T. M. Ellison [1994]: One-level Phonology: Autosegmental Representations and Rules as Finite Automata, *Computational Linguistics*, **20**, 1, pp. 55-90.

J-Ph. Bouchaud, M. Potters [1997]: Thorie des Risques Financiers. Portefeuilles, options et risques majeurs, Collection Ala Saclay.

N. Carels, P. Hatey, K. Jabbari, G. Bernardi [1998]: Compositional Properties of Homologous Coding Sequences from Plants, *Journal of Molecular Evolution*, vol. 46, pp. 45-53.

N. Chomsky [1957]: Syntactic Structres, The Hague: Mouton. Magyarul megjelent: Mondattani szerkezetek, Osiris-Szzadvg, Budapest, 1995.

N. Chomsky [1968]: Language and Mind, New York, etc., Harcourt, Brace & World, Inc. Magyarul megjelent: Nyelv s Elme, Osiris-Szzadvg, Budapest, 1995.

O. Clay, S. Cacciò, S. Zoubak, D. Mouchiroud and G. Bernardi [1996]: Human Coding and Noncoding DNA: Compositional Correlations, *Mol. Phylogen. Evol.*, Vol. 5., No. 1., pp. 2-12.

A. Czirk, R. N. Mantegna, S. Havlin, H. E. Stanley [1995]: Correlations in binary sequences and a generalized Zipf analysis, *Physical Review E*, **52**, 1 , pp. 446-452.

A. Czirk, H. E. Stanley, T. Vicsek [1996]: Possible origin of power-law behavior in  $n$ -tuple Zipf analysis, *Physical Review E* **53**, 6371.

M. Damashek [1995]: Gauging Similarity with  $n$ -Grams: Language-Independent Categorization of Text, *Science*, **267**, pp. 843-848.

I. Dernyi, T. Vicsek [1996]: The kinesin walk: A dynamic model with elastically coupled heads, *Proc. Natl. Acad. Sci. USA*, **93**, pp. 6775-6779.

G. Dietler, Y.-C. Zhang [1994]: Crossover from White Noise to Long Range Correlated Noise in DNA Sequences and Writings, *Fractals*, **2**, 4, pp. 473-479.

W. Ebeling, A. Neiman [1995]: Long-range correlations between letters and sentences in texts, *Physica A* 215, pp. 233-241.

F. Family, T. Vicsek [1991] (szerk.): *Dynamics of Fractal Surfaces*, World Scientific.

J. W. Fickett [1996]: The Gene Identification Problem: an Overview for Developers, to appear in the Proceedings of the Fourth International Workshop on Open Problems in Computational Biology, ed.: A. Konopka (as *Computers and Chemistry* 20, 103-118, 1996).

S. A. H. Geritz, J. A. J. Metz, . Kisdi, G. Meszna [1997]: Dynamics of Adaptation and Evolutionary Branching, *Physical Review Letters*, **78**, 10, pp. 2024-2027.

S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, Y. Dodge [1996]: Turbulent cascades in foreign exchange markets, *Nature*, **381**, 27 June 1996, pp. 767-770.

S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner [1997]: Turbulence and Financial Market, A Transaction Cascade in Foreign Exchange Markets, az 1997. júliusban, Budapesten tartott Econophysics Workshop során osztott preprint.

I. Große, H. Herzel, S. V. Buldyrev, H. E. Stanley [1997]: Mutual information of coding and noncoding DNA (preprint).

H. Herzel, W. Ebeling, I. Große, A. O. Schmitt [1997a]: Statistical Analysis of DNA sequences (preprint).

H. Herzel, I. Große [1995]: Measuring Correlations in Symbol Sequences, *Physica A*, **216**, 518.

H. Herzel, I. Große [1997b]: Correlations in DNA sequences: The role of protein coding segments, *Physical Review E*, **55**, 1, pp. 800-810.

I. Kanter, D. A. Kessler [1994]: Markov Process: Linguistics, Zipf's Law and Long-Range Correlations. (Submitted to PRL, Oct. 20, 1994).

R. M. Kaplan, M. Kay [1994]: Regular Models of Phonological Rule Systems, Computational Linguistics, **20**, 3, pp. 331-378.

Kiefer F. [1994] (szerk.): Strukturális magyar nyelvtan, 2. kötet: Fonológia, Akadémiai Kiadó, Budapest.

B. Krenn, Ch. Samuelsson [1996]: The Linguist's Guide to Statistics, forrás: <http://coli.uni-sb.de/~christer>.

Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, H. E. Stanley [1997]: Correlations in Economic Time Series (preprint submitted to Elsevier Science).

K. Mainzer [1997]: Thinking in Complexity, The Complex Dynamics of Matter, Mind, and Mankind, Springer, Berlin, etc.

Martins K. [1997]: A fenntarthat fejlődés termodinamikája (szemlyes példny)

K. Martins, M. Moreau [1995] (szerk.): Complex Systems in Natural and Economic Sciences, Proceedings of the Workshop Methods of Non-Equilibrium Processes...

R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H. E. Stanley [1994]: Linguistic Features of Noncoding Sequences, Physical Review Letters, **73**, 23, pp. 3169-3172.

R. N. Mantegna, H. E. Stanley [1995a]: Scaling behaviour in the dynamics of an economic index, Nature, **376**, 6 July 1995, pp. 46-49.

R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H. E. Stanley [1995b]: Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics, Phys. Rev. E, **52**, 3, pp. 2939-2950.

R. N. Mantegna, H. E. Stanley [1996]: Turbulence and financial markets, Nature, **383**, 17. October 1996, pp. 587-588.

R. N. Mantegna, H. E. Stanley [1997]: Stock market dynamics and turbulence, Parallel analysis of fluctuation phenomena, Physica A, 3357.

G. Matassi, L. M. Montero, J. Salinas, G. Bernardi [1989]: The Isochore Organization and Compositional Distribution of Homologous Coding Sequences in the Nuclear Genome of Plants, Nucleic Acids Research, Vol. 17., No. 13, pp. 5273-5290.

L. M. Montero, J. Salinas, G. Matassi, G. Bernardi [1990]: Gene Distribution and Isochore Organization in the Nuclear Genome of Plants, Nucleic Acids Research, Vol. 18, No. 7., pp. 1859-1867.

Nagy Ferenc [1986]: Kvantitatív Nyelvszet, Tankönyvkiadó, Budapest.

C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley [1992]: Long-range correlations in nucleotide sequences, *Nature*, **356**, pp. 168-170.

C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, H. E. Stanley [1993]: Finite-size effects on long-range correlations: Implications for analyzing DNA sequences, *Physical Review E*, **47**, 5, pp. 3730-3733.

M. Potters, R. Cont, J-Ph. Bouchaud [1997]: Financial markets as adaptive systems (preprint).

W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling [1989]: *Numerical Recipes in C, The Art of Scientific Computing*, Cambridge University Press.

I. Prigogine, I. Stengers [1986]: *La nouvelle alliance, Mtamorphose de la science*, Gallimard, Paris, 1986, *magyarul: Az j szvetsg, A tudomny metamorfzisa*, Akadmiiai Kiad, Budapest, 1995.

L. R. Rabiner [1989]: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, in: *Readings in Speech Recognition*, ed.: A. Waibel & Kai-Fu Lee, Morgan Kaufmann, 1991.

Rend s Kosz [1997], *Fraktlok s koszelmllet a trsadalomkutatsban*, szerk.: Fokasz Nikosz, Replika Kr, Budapest

Rvsz Gy. [1979]: *Bevezets a formlis nyelvek elmletbe*, Akadmiiai Kiad, Budapest.

J. Salinas, G. Matassi, L. M. Montero, G. Bernardi [1988]: Compositional Compartmentalization and Compositional Patterns in the Nuclear Genomes of Plants, *Nucleic Acids Research*, vol. 16, 4269-4285.

A. Schenkel, J. Zhang, Y.-C. Zhang [1993]: Long Range Correlation in Human Writings, *Fractals*, **1**, 1, pp. 47-57.

C. E. Shannon, W. Weaver [1986]: *A Kommunikci matematikai elmlete, az informcielmlet szletese s tvlatai*, OMIKK, Budapest.

M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, Ph. Maass, M. A. Salinger, H. E. Stanley [1996a]: Scaling behaviour in the growth of companies, *Nature*, **379**, 29 Febr. 1996, pp. 804-806.

M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, Ph. Maass, M. A. Salinger, H. E. Stanley [1996b]: Can Statistical Physics Contribute to the Science of Economics?, *Fractals*, **4**, 3 (1996), pp. 415-425.

H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, J. M. Hausdorff, S. Havlin, J. Mietus, C.-K. Peng, F. Sciortino, M. Simons [1992]: Fractal landscapes in biological systems: Long-range correlations in DNA and interbeat heart intervals, *Physica A*, 191, 1-12.

H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons [1993a]: Long-range power-law correlations in condensed matter physics and biophysics, *Physica A* 200, 4-24.

H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, S. M. Ossadnik, C.-K. Peng, M. Simons [1993b]: Fractal Landscapes in Biological Systems, *Fractals*, 1, 3, pp. 283-301.

Szpfalusy P., Tl T. [1982]: *A kosz*, Akademiai Kiad, Budapest

T. Vicsek [1992]: *Fractal Growth Phenomena* (second edition), World Scientific.

T. Vicsek, A. Czirk, E. Ben-Jacob, I. Cohen, O. Shochet [1995]: Novel Type of Phase Transition in a System of Self-Driven Particles, *Physical Review Letters*, vol. 75, 6, 1226-1229.

R. F. Weaver, Ph. W. Hedrick [1992]: *Genetics*, second edition, Wm. C. Brown Publishers

B. S. Weir [1990]: *Genetic Data Analysis, Methods for Discrete Population Genetic Data*, Sinauer Associates, Inc. Publishers, Sunderland, Mass.

G. K. Zipf [1935]: *The Psychobiology of Language*, Houghton Mifflin, Boston.

G. K. Zipf [1949]: *Human Behavior and the Principle of Least Effort*, Addison-Wesley Press.