0

DNS-SZEKVENCIÁK ANALÍZISE SZÖVEGELEMZÉSI MÓDSZEREKKEL

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

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

Diplomamunka

Budapest, 1998.

Összefoglal

Az elmlt nhny vtizedben mind a genetikban, mind a nyelvszetben hatalmas fejlőds kvetkezett be, amely lehetőv teszi azt, hogy a fizika, azon bell is a modern statisztikus fizika, j szempontok s kutatsi irnyok felvetsvel jruljon hozz e tudomnyok fejlődshez. A rohamosan nvekvő szmban hozzfrhető DNS-szekvencik, gnek, sőt egyre tbb egsz kromoszma s komplett genom, nem csupn lehetőv teszi, de − mint azt a msodik fejezetben megmutatom − ignyli is azt, hogy statisztikai eszkzkkel elemezzk őket.

Dolgozatom harmadik fejezetben ismertetni fogok tbb modern megkzeltst is (hossz tv korrelcik, Zipf-analzis, informcielmleti entrpia s redundancia), amelyeket termszetes nyelven rott szvegek s DNS-szekvencik analizlsra egyarnt alkalmaztak. E megkzeltsek arra az eredmnyre vezetnek, hogy a nem kdol DNS-szakaszok sok tulajdonsgukban hasonltanak az rott szvegekre. Mivel kiderl, hogy Markov-folyamatokkal nem rhatk le kellő pontossggal ezek a statisztikus tulajdonsgok, a sztochasztikus krnyezetfggetlen grammatikk elmlett fogom javasolni e jelensgek lersra. Vgl diplomamunkm negyedik fejezetben egy vektortrtechnika alkalmazst mutatom be DNS-szekvencikra. A mdszer alkalmasnak bizonyult rott szvegek nyelv s tartalom szerinti sztvlogatsban, most a kdol s a nem kdol DNS-szakaszok elklntsre is sikerrel alkalmazzuk.

Szeretnm ezen a ponton ksznetemet kifejezni tmavezetőmnek, Vicsek Tams tanszkvezető egyetemi tanrnak, aki minden tren nagyban hozzjrult e dolgozat elkszlthez. Ksznet illeti Major gnest is, aki az ELTE TTK Genetika tanszke rszről a biolgus szemszgből segtette a munkmat, valamint Czirk Andrs doktorandusznak, akitől a Damashek-mdszer DNS-szekvencikra trtnő alkalmazsnak az tlete szrmazik, valamint sok tovbbi rtkes tanccsal s szakirodalommal segtett.

Tartalomjegyzk

1. Bevezets **4**

1.1. A statisztikus fizika interdiszciplnris alkalmazsaibl 6
1.2. Biolgia s statisztikus fizika 8

2. A DNS-szekvencik kutatsnak irnyai **11**

2.1. A statisztikai megkzelts kt clja 11
2.2. A gnidentifikci mdszerei 14

3. DNS-szekvencik s rott szvegek modellezse statisztikai eszkzkkel **17**

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 generatv 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

4. A vektortr-technika **42**

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

5. sszefoglals **58**

Irodalomjegyzk **61**

1. fejezet

Bevezets

"A clash of doctrines is not a disaster, it is an opportunity."
Whitehead [1]

gy tűnik, a msodik vilghbort, ill. az azt megelőző, politikailag feszlt lgkrt s az jjpts gazdasgi nehzsgeit kvetően, az 1950-es vekben j lendletet vehetett a tudomny. J nhny diszciplnban forradalmi vltozsok zajlottak le ezekben az vekben. A szmtstechnika hőskorszaka s az űrkutats kezdete (1957) egybeesik az első atomerőmű (1954) felpltvel, az antiproton (1955) s a paritssrts (1956-57) felfedezsvel. 1956-ban vezeti be Rindler az esemnyhorizont fogalmt. s hogy teljesen ms terletről is idzznk pldkat, az tvenes vek vgn kezd kibontakozni a kognitv pszicholgia, valamint az vtized kzepn publikljk az 1947-től kezdve felfedezett Holt-tengeri tekercsek első fragmentumait, amelyek a Biblia-tudomnyban, az kori Palesztina trtnetnek a kutatsban s a hber nyelvszetben egyarnt vratlan fordulatot hoztak.

Bennnket most hrom, első pillantsra nagyon tvolinak tűnő tudomny ezekben az vekben indult megjulsa fog rdekelni. Időrendben haladva, elsőnek a *genetika* forradalma zajlott le: 1953-ban jelentette meg *J. D. Watson* s *F. H. C. Crick* azt a rvid cikket, amely nlkl nem kpzelhetnnk el az elmlt negyven v biolgijt. *A. N. Kolmogorov* 1954-es ttele (*V. I. Arnold* s *J. K. Moser* ltalnostsaival, az n. *KAM-ttel*), a statisztikus fizikban nyitott j tvlatokat, amely a kosz-elmlethez, a komplex rendszerek fizikjhoz, a fraktlok fizikai alkalmazsaihoz vezetett (Szpfalusy & Tl [1982], Mainzer [1997]). Vgezetl, 1957-ben ltott napvilgot *N. Chomsky Syntactic Structures* cmű műve, amely − fggetlenl attl, hogy egyetrtnk-e az előfeltevseivel

[1] *"Ha a doktrnk tkznek egymssal, az nem katasztrfa, hanem kihasznland esly.", Whitehead, Science and Modern World, p. 186, idzi Prigogine & Stengers [1986].*

– a nyelvtudomny teljesen j megkzeltsi mdjt jelentette, a korbbi, strukturalista nyelvszettel szemben.

Mi hozza kzs nevezőre ezt a hrom tudomnygat? Mindhrom tudomnyban olyan krdseket lehet a szzad kzeptől kezdődően egzakt formban feltenni s a termszettudomnyos modellalkot-jslatot ellenőrző kutatsi mdszer segtsgvel megvlaszolni, amelyek a megelőző vszzadokban csupn a filozfiai spekulcik trgyai lehettek.

A modern genetika vglegesen vget vetett tbb ezer v legklnflbb dlibbos elkpzelseinek: az ősi elkpzelsnek, amely a vrhez kapcsolta az rklődst,[2] valamint sok korai "tudomnyos", de nem megfigyelsre alapozott elkpzelsnek (Berend M. [1980]), amelyeket rendre megcfoltak. Ilyen volt pldul az *ősnemzs tana*, mely az isteni lleknek (ugyanis ezen elkpzels szerint a nemzs legfontosabb pillanata "*a llek beltetse a test anyagaiba*") vagy a hmmagnak tulajdontotta az utd tulajdonsgait. *Paracelsus* (1493-1514) svjci termszettuds s orvos tanai, amelyek szerint a terhes nő vgyai, lmai, benyomsai jelennek meg a szletendő gyermek tulajdonsgaiban, mig hatnak a szlesebb trsadalmi rtegek gondolkodsmdjban. A XVII-XVIII. szzad *preformizmustannak* hirdetői szerint pedig az ivarsejtekben *preformlt homunkuluszok*, miniatűr emberkk tallhatak, amelyeknek szintn vannak mg kisebb homunkuluszokat tartalmaz ivarsejtjei, s gy tovbb. A genetika tudomnya akkor jtt ltre, amikor a ksrletek mr nem utlag cfoltk meg a spekulatv elmleteket, hanem a ksrletek eredmnyeiből kiindulva ptettk fel őket. *G. J. Mendel* (1822-1884) trvnyei, majd a XX. szzad első felnek kutatsai (*H. de Vries*, *C. E. Correns*, *A. v. S. Tschermak*, *A. Weismann*, *O. T. Avery*, *T. H. Morgan* s sokan msok) tettk lehetőv, hogy *Watson* s *Crick* felfedezse megalapozza az rklőds igazi elmlett.

Noam Chomsky kvetői immron szinte termszettudomnynak tekintik a nyelvszetet. A korbbi szemllet szerint e tudomny clja a nyelvi rendszer lersa volt, s ezt – elmleti megalapozottsgban, precizitsban – a strukturalistk vittk vgbe a legtkletesebb mdon, a XX. szzad első felben. Chomsky első műve utn, akr kvetjk iskoljnak szintn vltoz, fejlődő gondolatait, akr nem, a nyelvsz feladata egy jval mlyebb lerst ad s minl adekvtabb *modell* alkotsa, amelyek eleget tesznek bizonyos, a modellel szemben tmasztott, elmleti elvrsoknak is.[3] Az első modellek, akrcsak a Newton-i mechanika termszetmodellje, csupn egy nagyon szűk s idealizlt jelensgkr lersra voltak alkalmasak, de a tudomny lass fejlődse sorn szlesedett (s, remljk, szlesedni fog) a lerhat adathalmaz.

[2] Gondoljunk pldul a "kkvrű", "vrvonal" vagy a "vrfertőzs" kifejezsekre.

[3] A modellnek egysges lerst kell adnia minl tbb nyelv minl vltozatosabb jelensgeiről, meg kell felelnie nhny nyelvszeti elvnek, figyelemmel kell lenni a gyermek nyelvelsajttsnl s affzis betegeknl megfigyelt jelensgekre, s pszicholgiailag is adekvtnak kell lennie.

A chomskynus nyelvszet msik nagy eredmnye a matematikai eszkzk bevonsa a modellekbe. Idzzk Chomsky szavait:[4] "*Figyeljk meg, hogy egybknt az egyetemes nyelvtan pontosan krlrt elveinek ltezse teszi lehetőv egy j terlet, a matematikai nyelvszet ltrejttt, mely az egyetemes nyelvtanban meghatrozott megszortsoknak eleget tevő generatv rendszerek osztlyt veti elmleti vizsglat al. Ez a vizsglat valamennyi lehetsges emberi nyelv formlis sajtsgainak rszletes feltrsra trekszik. A terlet mg gyerekcipőben jr; csak az utbbi vtizedben vlt egy ilyen vllalkozs lehetősge elkpzelhetőv. ... Úgy tűnik teht, hogy a matematikai nyelvszet jelen pillanatban egyedlllan kedvező helyzetben van a trsadalomtudomnyok s a pszicholgia matematikai megkzeltsei kztt ahhoz, hogy ne egyszerűen adatok elmletv,*[5] *hanem az emberi mentlis folyamatok jellegt meghatroz rendkvl elvont elvek s struktrk vizsglatv fejlődjn.*"

Ezek a matematikai nyelvszeti modellek nem csupn a kognitv tudomnyokra vannak jelentős hatssal, nem csupn egzakt formba ntttk a korbbi fogalmakat, s nem is csak a nyelvtudomny tudomnyossgi kritriumait vltoztattk meg, hanem potencilisan lehetősget teremtenek arra, hogy a nyelv kvantitatv viszonyait is beptsk a nyelvszeti modellekbe. Ezen a ponton kapcsoldik a nyelvszethez a fizika, mint eredetileg az (lettelen?) termszet kvantitatv jelensgeit ler tudomny, s dolgozatom kső̋bbi rszeiben be fogom mutatni eme vizsglatok első, főleg fizikusok ltal produklt eredmnyeit is.

1.1 A statisztikus fizika interdiszciplnris alkalmazsaibl

A genetikban s a nyelvszetben lezajlott vltozsok utn, most trjnk vissza az 1950-es vek nagy, tudomnytrtneti fordulataihoz, azon bell is a fizika j gaihoz. A modern statisztikus fizika tette lehetőv az nszerveződő komplex struktrk vizsglatt. A klasszikus mechanika, az időnyl megfordthatsga miatt, csupn a legegyszerűbb folyamatokat, pldul a bolygmozgst vagy az ingamozgst volt kpes lerni. A mlt szzadi termodinamika irreverzibilis vilga les ellentmondsba kerlt a reverzibilis mechanikval, de tovbbra is csak a homogn egyensly, a "hőhall" irnyba trtnő vltozsokat tudta rtelmezni. Csupn az elmlt vtizedek fejlődse eredmnyeknt fogalmazdott meg az, hogy "*az irreverzibilitsnak a termszetben alkot szerepe van, hiszen spontn szerveződsi folyamatokat tesz lehetőv. A fizikn bell az irreverzibilis folyamatok tudomnya visszalltotta jogaiba a tevkenysg- s nszaport struktrkat ltrehoz termszet*

[4] Chomsky [1968], magy. kiads 227. o.

[5] *Chomsky itt a kvantitatv nyelvszetre utal (ld. pl. Nagy F. [1986]-t), amely egyszerű statisztikai megfigyelsekből von le nem igazn messze mutat, helyenknt erősen vitathat kvetkeztetseket. (Itt pldul a glottokronolgia kritikjra gondolok.)*

gondolatt." (Prigogine-Stengers [1986] p. 10.) Megnylt ezzel az t az sszetett rendszerek fizikai eszkzkkel trtnő vizsglathoz, ahhoz, hogy az alkot elemektől fggetlenl, a struktrt is lehessen elemezni.

Ugyanakkor a termszetben előfordul sszetett rendszerek vizsglata, ppen a fizika korbbi gainak emltett jellegből addan, kiesett e tudomny hagyomnyos trgykrből. Ezrt okoz sokaknak meglepetst, ha pldul biolgiai, kzgazdasgi vagy szociolgiai rendszerek vizsglatba fizikusok is bekapcsoldnak. Pedig eme interdiszciplinris kutatsok manapsg egyre tbb fizikust vonzanak, s nagyon gretesnek tűnnek.

A nem-lineris statisztikus fizika fedezte fel az 1980-as vekben azt, hogy egy sor termszeti jelensg fraktlszerű kpződmnyeket hoz ltre a termszetben (Family & Vicsek [1991], Vicsek [1992]). Ezt kvetően, a fraktlok s a kaotikus jelensgek kutatsnak "divatja" tterjedt a trsadalomtudomnyokba, sőt a művszetelmletbe is. A *Rend s Kosz* [1997] cmű, a kzelmltban megjelent tanulmnyktetben a koszelmlet trtnelmi s kzgazdasgtani alkalmazsai mellett, kpzőművszeti s zenei alkotsok fraktltulajdonsgokkal trtnő elemzsre is tallunk ksrleteket.

A kzgazdasgtanban egy sor jelensg elemzsnl merlhetnek fel a statisztikus fizikbl klcsnztt fogalmak. (Tovbbi pldkat hoz Mainzer [1997] pp. 270-281. is.) Itt nem is elsősorban a termodinamikai fogalmak s trvnyszerűsgek gazdasgi kvetkezmnyeinek a vizsglatra gondolok, amelyre pldt szolgltat *Csekő . & K. Martins* (in: K. Martins, M. Moreau [1995]), valamint Martins [1997]. Az 1990-es vek kzepn (jbl) a statisztikus fizikval foglalkoz kutatk rdeklődsi krbe kerltek egyes sklzsi tulajdonsgot mutat kzgazdasgi jelensgek s kaotikusnak tűnő folyamatok.

Az egyeslt llamokbeli ipari cgek msfl vtizedet tfog adatai alapjn azt tapasztaltk, hogy az azonos ves forgalommal rendelkező cgek kztt az ves logaritmikus nvekedsi rta exponencilis eloszlst mutat (M. H. R. Stanley et al. [1996a,b]):

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

ahol s_0 az ves S_0 forgalom logaritmusa egy adott vben, r pedig ennek vltozsa a kvetkező vben ($r = \ln(S_1/S_0)$). Az eloszls $\sigma(s_0)$ szrsa cskken nvekvő forgalom mellett, mghozz meglepő sklzsi tulajdonsggal rendelkezik:

$$\sigma(s_0) = a\exp(-\beta s_0) = aS_0^{-\beta}. \tag{1.2}$$

A sklzsi exponens $\beta = 0,15 \pm 0,03$, ha az eladsok vltozst vizsgljuk, ill. $\beta = 0,16 \pm 0,03$, ha az alkalmazottak szmt tekintjk a fentiekkel analg mdon. A jelensg magyarzatra izgalmas modellel szolgl Amaral et. al. [1997].

Gazdasgi mutatk (az n. Standard & Poor's 500 index) valsznűsgi eloszlsnak sklzsi tulajdonsgt mutatta meg Mantegna & Stanley [1995a], hrom nagysgrenden keresztl, az 1 perctől az 1000 percig tart időskln (ld. mg Y. Liu et al. [1997]).

A rvid tv, spekulatv valutapiac jelensgeit Ghashghaie et. al. [1996, 1997] (ld. mg Mantegna & Stanley [1996, 1997]) a hidrodinamikai turbulencikhoz hasonltja, mivel az előbbiben olyan "informci-kaszkdokat" ttelez fel, amelyek a hidrodinamikai turbulencik energia-kaszkdjainak felelnek meg.

Jean-Philippe Bouchaud s trsainak munkssga az elmlt vekben a pnzgyi kockzatok s az opcik razsnak terletn hozott j eredmnyeket (Bouchaud & Potters [1997], Potters et al. [1997]): az opcik razsnl hasznlatos, 1973-bl szrmaz s 1997-ben kzgazdasgi Nobel-djjal jutalmazott *Black − Sholes-formula* hinyossgaira mutatnak r, mind elmleti szmtsokkal, mind valdi adatok feldolgozsval.

A statisztikus fizika kzgazdasgtani s pnzgyi alkalmazsaibl hozott felsorols vgn hadd emltsnk egy magyar pldt is: Vzvri Bla s Bacsi Zsuzsanna (*in*: Rend s Kosz [1997]) a magyar burgonyapiac viselkedsben figyelt meg kaotikus tartomnyokat. Miutn az egytermkes piacmodellt illesztettk a megfigyelt adatokra, megvizsgltk a modell reaglst az intervencis rak vltoztatsra: itt is megfigyeltk a nem-lineris dinamikai rendszereknl megszokott periodikus s kaotikus tartomnyok megjelenst, ill. peridus-tbbszrző bifurkcikat, a paramterek vltoztatsa sorn.

1.2 Biolgia s statisztikus fizika

ttrve a termszettudomnyokra, nagyon nagy szmban vannak a fraktlis viselkedst mutat termszeti jelensgek (Family & Vicsek [1991], Vicsek [1992]), a kristly-nvekedstől kezdve, a domborzat s a tengerpart vonalnak alakulsig. Kaotikus jelensgek megigyelhetők szmos kmiai reakciban (Szpfalusy & Tl [1982]).[6]

Önszerveződő folyamatokra az lővilgbl merthetjk a legjobb pldkat. Az egyik első vizsglt plda a *Dictyostelium discoideum* nevű amőbafaj sejtpopulciinak ltvnyos talakulsa (Prigogine & Stengers [1986], pp. 148-149): ha a telep felemsztette krnyezetben a tpanyagot, a sejtek sszellnak egy masszv, amelyben differencici tjn ltrejn egy szr, azon massza, s ebből sprk radnak ki j, tpanyagban gazdag lőhelyek fel. Ez a folyamat "spontn mdon" jn ltre, az amőbapopulciban vndorlsi hullmok alakulnak ki, egy − a vletlen fluktuci ltal ltrehozott −

[6] A brsszelltor s a Bjeluszov−Zsabotyinszkij-reakcival kapcsolatban ld. mg Prigogine & Stengers [1986]-t (pp. 149-153), amely a kvetkező cikkre hivatkozik: A. Winfree: Rotating Chemical Reactions, Scientific American, vol. 230 (1974), 82-95.

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

Hasonl komplex struktrk spontn ltrejttt figyelte meg Vicsek et al. [1995] "nhajt" rszecskkből ll rendszerben: habr az egyes rszecskk, amelyek pldul egy madrraj tagjait modellezhetik, nllan mozognak, bizonyos kritikus sűrűsg fltt s zaj alatt fzistalakuls trtnik a rendszerben, a mozgsirnyok egysges rendbe rendeződnek. Ennek oka a kzeli szomszdok mozgsirnynak tlagbl definilt rvid tv klcsnhats, amely mgis az egsz 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 klnbző evolcis stratgik elterjedst vagy kihalst vizsgljk, ha j mutcik, azaz j evolcis stratgik jelennek meg. "Elgazsokat" figyeltek meg, azaz egy faj kettvlst kt fajj. A kutatk clja az, hogy a modell idealizltsgnak cskkentsvel az evolci egyre tbb jelensgt legyenek kpesek lerni.

Egszsges emberek kt szvverse kztt eltelt időtartam időbeli vltozsait vizsglta 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 viselkedsből ($C(t) \sim t^{\gamma}$, $S(f) \sim f^{\beta}$) 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 magyarzhatk lennnek, a hossz tv korrelcik ltre a szerzők nem adnak magyarzatot.

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

Dolgozatom trgya, a DNS-szekvencik elemzse, a statisztikus fizika azon fejezeteihez tartozik, amelyek *szimblumsorozatok*, gy pldul termszetes nyelven rott szvegek vagy szmtgpes programok, statisztikai vizsglathoz 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 kutatsnak irnyai

2.1 A statisztikai megkzelts kt clja

Az elmlt vekben, msfl vtizedben exponencilisan nőtt az ismert DNS-szekvencik mennyisge. Ma mr szinte automatikuss s automatizlhatv vlt az a munka, ami vtizedekkel ezelőtt mg hossz, vertkes s sok emberi lelemnyt ignylő feladatnak szmtott. A 2.1. tblzat s a 2.1. bra az *NCBI-Genbank* (megtallhat a `http://www.ncbi.nlm.nih.gov` cmen) adatbzisban trolt adatmennyisg időbeli vltozst mutatja.

A szekvencik feldertsnek automatizlhatsga vetette fel azt az ignyt, hogy a szekvencik "jelentsnek" meghatrozst is automatizlni kell ahhoz, hogy a kt munkafzis tempt tudjon tartani egymssal. A problmt itt nem is a kdol rszek aminosav-szekvenciv trtnő lefordtsa jelenti, hiszen a kdsztrat az 1960-as vekben megfejtettk (Berend [1980], 41. o.). A feladat a kdol s a nem kdol rszek klnvlasztsa, a gnek identifiklsa, s ez a krds mg messze van attl, hogy teljesen megoldottnak tekinthessk, klnsen a magasabb rendű eukariota lőlnyek esetben (Fickett [1996]).

Fickett szerint a cl az "automatikus annotci", azaz a szekvencia egyes szakaszaihoz hozzrendelni azok "tulajdonsgait", minl teljesebb s pontosabb mdon. Ezek a "tulajdonsgok" sokflk lehetnek, de a legjelentősebb krds az egyes szakaszok *funkcionlis* szerepnek (fehrje kdolsa, szablyozs,...) a meghatrozsa. Ezen eredmnyek egyik nagyon fontos clja Fickett szerint − hossz tvon − olyan ltalnostsok levonsa, amelyek szablyokat adhatnak a gneket s genomokat tervező, jvőbeli kutatk kezbe. Jelenleg viszont mg meg kell elgednnk azzal, ha az algoritmusunk elfogadhat pontossggal jsolja meg, mely szekvencia kdol fehrjt. Ha a fehrje funkcijra is kapunk tippet, gy az mr jelentős eredmny. Az aminosav-szekvencit

Release	Dtum	Bzisok szma	Ttelek szma
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 informcimennyisg (ttelek szma, ill. az azokban sszesen lvő bzisok szma) időbeli nvekedse. A GenBank-ben tallhat bzisok szma krlbell 14 hnaponknt ktszereződtt meg. (Forrs: `http://www.ncbi.nlm.nih.gov`*)*

kzvetlenl nem kdol DNS-szekvencik automatikus csoportostsa, annotlsa mg nagyon tvolinak tűnik, ugyanis, mint azt ltni fogjuk, ezek jelentős rsznek szerepről mg nincs sok elkpzelsnk. Ehhez kapcsoldik a msodik problmakr, amely a DNS-szekvencik kutatsnak irnyt meghatrozza.

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

2.1. bra: Az NCBI-GenBank fejlődsi teme, amely jl kzelthető exponencilissal, s krlbell 14 havonta ktszereződik meg az ismert szekvencik bzisainak sszege. (Forrs: `http://www.ncbi.nlm.nih.gov`*)*

& Hedrick [1992]). Ezek egy rsze mra mr jl ismert. Pldul a *TATA-box* 25 darab timin s adenin vltakoz sorozatbl ll, s a legtbb eukarita sejt promotereinek jellegzetes rsze, amely az RNS-polimerz bektődst, s ez ltal a transzkripci megkezdst teszi lehetőv. Ms szekvencik szintn jelentős szerepet jtszanak a transzkripci szablyozsban.

Viszont sok szekvencia szerepe mg ismeretlen. Nmely kutatk hajlanak arra, hogy ezek egy rszt "evolcis szemtnek" tekintsk. Szerintk olyan szakaszokrl van sz, amelyekről megszűnt a transzkripci az evolci sorn, elvesztettk jelentősgket. Így szabadon, fenotipikus kvetkezmnyek nlkl, ezrt knnyen mutldhatnak, vagyis az eredeti "jelents" (pldul kdolt aminosav-szekvencia, szablyoz szerep) mr nem felismerhető.

Van olyan vlemny, amely szerint az intronok szerepe abban ll, hogy mg *redundnsabb* tegyk a genetikai kdot, tovbb cskkentve ezltal a mutcik veszlyt. Kzismert, hogy mr a kdsztr is tartalmaz bizonyos redundancit (a genetikai kd *degenerlt*). Hiszen bzis-hrmasok (triplettek, kodonok) kdoljk az aminosavakat, mrpedig a ngy bzis 64-fle triplettet alkothat, s mg ha le is vonjuk a *stop*-jelnek megfelelő hrom kodont, 61 bzis-hrmas tartoztik a 20 (21) aminosavhoz. Ugyanakkor, sokak szerint a nem kdol szakaszok a proteinek elsődleges struktrjn tlmutat, biolgiailag fontos informcit tartalmazhatnak (Mantegna et al. [1995b]).

Klnsen "rejtlyes" az *intronok* szerepe. Ezek olyan szakaszok, amelyek az aminosav-szekvencit kdol *exonokkal* egytt trdnak a DNS-ről az mRNS-re (*primary transcript*, azaz elsődleges tirat), viszont később valamilyen mechanizmus kivgja az intronokat az mRNS-ből, gy azok nem kerlnek bele a proteint szintetizl folyamatokba. Egy magasabb rendű faj primary transcript-je tipikusan tartalmaz egy *vezrszakaszt*, amelynek a szablyozsban van szerepe, egy *vgszakaszt*, exonokat, valamint az exonokat egymstl elvlaszt intronokat.

A gneken belli s a gnek kztti nem kdol szakaszok sszetettebbekk teszik a gneket, genomokat. A 31 nukleotidbl ll virlis SV40 gn s a 210 000 bzist tartalmaz 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 semmifle 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-kopoltys 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 analzisből arra a kvetkeztetsre jutottak a kutatk, hogy a kdol s a nem kdol szakaszok arnya *cskken* az egyszerűbb fajoktl a magasabb rendű fajok fel. Pldaknt, a Mantegna et al. [1995b] cikkben elemzett emlős (ember, egr) szekvencik esetben $5,3\%$, az lesztő kt kromoszmja esetben 70% krli, mg az *E. coli* baktrium hat szekvencija s kt fg teljes genomja esetben a 80%-ot is meghaladja a kdol szakaszok arnya az egsz szekvencihoz kpest. Ez a megfigyels magyarzhat mind az "evolcis szemt"-hipotzissel, mind azzal az elkpzelssel, amely szerint ezeknek a szekvenciknak igenis van jelentősgk, pldul a transzkripci finomabb szablyozsban, s ezek a mechanizmusok fejlettebbek a magasabb rendű lőlnyekben. (Mantegna et al. [1995b])

2.2. A gnidentifikci mdszerei

A feladat teht a DNS klnbző tpus szakaszainak, olyan "tulajdonsgait" megtallni, amelyek alapjn sztvlaszthatk ezek a tpusok. Egy-egy ilyen tulajdonsg nem csupn egy DNS-t annotl automatikus algoritmus alapjul szolglhat, hanem hozzjrulhat a klnbző szakaszok szerepnek mlyebb megrtshez is. A kutatsok jelenlegi stdiumban elsősorban a kdol s a nem kdol szekvencik klnbzősgre, a *gnek identifiklsra* kell sszpontostani az erőinket. Dolgozatomban tovbb szűktem a krdst, az exonok s az intronok eltrő viselkedst vizsglom.

A mdszereket *Borodovsky*, *Koonin* s *Rudd* alapjn,[2] Ficket [1996] kt csoportra osztja. Az első csoport, a "mintzat-kereső mdszerek" (*template methods*, *intrinsic approaches*), a keresett objektum "prototpust" 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 jtszanak a transzkripciban, "*signal*", "ktőhely" (*bind-*

[1] A *bp* a *bzispr* (*base pair*) kifejezs rvidtse, s a genetikai irodalomban a DNS-szekvencik hossznak mrtkegysgeknt hasznljk. Hasznlatosak a *bp* tbbszrsei, a *kbp* s a *Mbp* is.

[2] 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, konkrt funkcit betltő szekvencit keres a genomban. De mivel egy "signal"-fajtnak is sok vltozata ltezik, először "konszenzus-szekvencia" definilsval prblkoztak: egy adott "signal" sszes előfordulst egyms mell helyeztk, s az idealizlt szekvencit gy kaptk meg, hogy mindegyik pozciba az ott leggyakrabban előfordul bzist vettk fel.

De ez a mdszer tlsgosan egyszerűnek s nem elg hatsosnak bizonyult, ezrt az n. *mtrix-mdszerhez* kellett folyamodni. Jellje $f(b,i)$ a b bzis előfordulsi gyakorisgt a signal klnfle vltozatainak i-ik pozcijban, $p(b)$ pedig a b bzis frekvencijt a genomban. Az

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

mtrixot nevezzk a "signal" "pozci-sly mtrixnak" (*position weight matrix*). Amikor egy szekvencirl el akarjuk dnteni, vajon az "signal"-szekvencia-e, kpeznnk kell a tesztelendő szekvencia pozci-sly mtrixt is:

$$s(b,i) := \begin{cases} 1, & \text{ha a szekvencia } i\text{-ik pozcijban a } b \text{ bzis szerepel} \\ 0 & \text{egybknt} \end{cases} . \tag{2.2}$$

A teszt eredmnyt, amely megadja, vajon mennyire "hasonlt" a tesztelt szekvencia a "signal"-ra, a kt mtrix skalris (elemenknti) szorzataknt kapjuk meg:

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

Ha ez az S rtk magas, nagy valsznűsggel lesz a vizsglt szakasz "signal". Az eljrs lert tovbbfejlesztse mr a gneket felismerő mdszerek msodik osztlynak technikai rszleteihez ll kzel.

A "mintzat-kereső mdszerekkel" szemben llnak, s utbbiak egyre nagyobb szerephez jutnak az ismert szekvencik szmnak rohamos nvekedsvel, a "visszakereső mdszerek" (*lookup* vagy *extrinsic methods*). Ezen algoritmusok az ismert szekvencik sszehasonltsbl vonnak le kvetkeztetseket. A legjobb pldt ez utbbi eljrscsoportra a szekvencik kztti hasonlsgot definil mdszerek adjk.

Kt, azonos hosszsg DNS-szekvencia kztti "tvolsg" legegyszerűbb mrtke gy hatrozhat meg, hogy pronknt sszehasonltjuk a kt szekvencit alkot bzisokat, s vesszk az eltrő bzisok szmt (Weir [1990]). De ezen mdszer, illetve tovbbfejlesztett vltozatai, csupn bizonyos clokra, mghozz a biolgusok szmra igen fontos, fajok kztti genetikai tvolsgok meghatrozsra

alkalmasak, amelyek az evolcis-genetikai rokonsgok meghatrozshoz szksgesek (*phylogeny construction*, *parsimony methods*).

Ennl finomabb, szlesebb krben hasznlhat mrtkek a szekvencik bizonyos statisztikai tulajdonsgait hasznljk fel. Ficket [1996] kzel hsz ilyen mdszert sorol fel, a kodonok, hexamerek (bzishatosok), aminosavak,[3] aminosavprok előfordulsi gyakorisgt kdol vektoroktl kezdve, a szekvencia Fourier-transzformltjait tartalmaz vektorig. Ezekkel a mdszerekkel a gneknek legalbb a felt lehet mr ma is automatikusan felismerni, s sok esetben a gn funkcijra is lehet kvetkeztetni.

Ezek a mdszerek a gnek azonostsra szolglnak. A dolgozatom fő irnynak az exonok s az intronok, ill. az intronokat tartalmaz s nem tartalmaz gnek kztti statisztikai klnbsgek vizsglatt vlasztottam. Ezen klnbsgek felismerse egyszerre jrulhat hozz mindkt, a 2. fejezet elejn felvetett problma megoldshoz: egyrszt alapjaiv vlhatnak olyan jabb automatikus algoritmusoknak, amelyek kpesek lesznek sztvlasztani a ktfle szekvencit, msrszt az intronok tulajdonsgainak megismerse fnyt vethet a jvőben az intronok szerepre is.

[3] Az aminosavak frekvencija Mantegna et al. [1995b] szerint csupn kis klnbsgeket mutat, ha nagyobb szekvencia-halmazokra tlagolva vizsgljuk. Ezzel szemben, a kodonok gyakorisgnak eloszlsban nagy eltrseket tallhatunk.

3. fejezet

DNS-szekvencik s rott szvegek modellezse statisztikai eszkzkkel

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 legtbbszr 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, szkzepi s szvgi helyzetben is.

Weir [1990] az egyes bzisok gyakorisgra hoz pldkat, klnbző lokuszokban.[1] Pldaknt, Weir 7.5 tblzatban (233. oldal) szerepel a karfiol mozaik-vrus, amely MCACGDH gnjben az adenin arnya 37%, mg a ΦX174 bakteriofg PX1CG lokuszban csupn 24%. Az egr mitokondriumban 12% a timin arnya, a Hepatitis B vrus HPBAYW gnjben viszont 27%.

A biolgusok a leggyakrabban a DNS-szekvencik n. *GC-arnyt* (vagy *G+C-arnyt*) vizsgljk. 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 vizsglja klnfle szervezetek genomjainak szerkezett, a GC-szint fggvnyben. Salinas et al. [1988] kimutatja, hogy egyes nvnyfajok genomja mozaikszerűen pl fel a GC-szint szempontjbl homogn 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]).

[1] A lokusz (*locus*) olyan gnt jell, amelynek ismert a pontos pozcija a kromoszmn.

Ezek a kutatsok egyttal az exonok s az intronok GC-szintjvel kapcsolatban is hoztak j eredmnyeket. A Salinas et al. [1988] ltal vizsglt nvnyekben a GC-szint jval alacsonyabb valamely gn intronjaiban, mint az adott gn exonjaiban, de a kt rtk kztt lineris korrelci fedezhető fel egyes fajok genomjn bell. Az is megfigyelhető, hogy azoknl a fajoknl, amelyeknl tlagosan alacsonyabb az exonok GC-szintje, alacsonyabb az intronok GC-szintje is. Hasonl eredmnyek szerepelnek, szintn nvnyek kapcsn, Carels et al. [1998] 6. brjn, valamint hasonl korrelcikrl szmol be az emberi genomban Clay et al. [1996]. Ezek a tnyek rszben indokolhatjk a 4. fejezetben ismertetsre kerlő sajt eredmnyeimet is.

A kdol szekvencikon bell, a triplettek (kodonok) eltrő pozciiban eltrő az egyes bzisok gyakorisga. Herzel & Große [1995], ill. Herzel et al. [1997a] az lesztő III. kromoszmjnak adatait, Herzel & Große [1997b] pedig az emberi β-miozin HUMBMYH7 gn rtkeit idzi: valamely bzisnak a kodon első, msodik, ill. harmadik pozcijban val előfordulsi gyakorisga kztt gyakran kt-hromszoros szorz is lehet.[2]

A klnbző pozciban mrhető GC-szintekre GC_1, GC_2, ill. GC_3-knt szoks hivatkozni. Mivel a harmadik pozci ltalban redundns (legalbb rszben[3]), ezrt rdekes a GC_{1+2}-szint is, amely a guanin s a citozin koncentrcijt jelenti az első kt pozciban. A klnbző pozcikban mrhető GC-szintek kztti lineris korrelcikrl szintn beszmolnak Bernardi csoportjnak idzett cikkei.

Tovbblpve az egyes bzisok frekvencijn, vizsglhatjuk nukleotid-kettesek gyakorisgt is. Weir [1990] erre is hoz pldt (ld. Weir 7.8. tblzatt). Megfigyelhető, hogy a szomszdos bzisok nem fggetlenek egymstl, vagyis ha p_i jelli az i bzis előfordulsi valsznűsgt, p_{ij} pedig az (ij) bzis-kettes gyakorisgt, akkor

$$p_{ij} \neq p_i p_j. \tag{3.1}$$

Erre szolgltat pldt Weir [1990] 7.7 tblzata.

3.1 Markov-folyamatok

A szomszdos bzisok kztti korrelci szolgltatta az tletet, hogy a DNS-szekvencikat *Markov-lncknt* lehetne modellezni (Weir [1990]). Az egyszerű statisztikkon tllpve, a sztochasztikus

[2] A legszlsősgesebb klnbsg a HUMBMYH7 gn msodik s harmadik kodon-pozcija kztt fordul elő: az adenin frekvencija a msodik pozciban $43,7\%$, mg a harmadik pozciban csupn $7,9\%$!

[3] Pldul mind a ngy `TC`-vel kezdődő kodon a *szerin* nevű aminosav megfelelője a kdsztrban; a `CAT` s a `CAC` *hisztidint* kdol, a `CAG` s a `CAA` pedig *glutamint* (Berend [1980]).

eszkzkkel trtnő modellezs j skra emeli a DNS-szekvencik tanulmnyozst, s ezen a ponton tallkozik a genetika a statisztikus fizika 1. fejezetben emltett interdiszciplnris alkalmazsaival.

Sztochasztikus folyamatnak egy olyan fizikai (vagy brmilyen 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 karakterszekvenciák, 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 ($\Omega = \{s_1, ..., s_n\}$), egy véges ábécé ($\Sigma = \{\sigma_1, ...\sigma_m\}$), 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 empírikus 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 Ω^* s a Σ^* az Ω, ill. a Σ 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 modellezse Markov-folyamatknt Weir [1990] szerint inkbb az egsz genom, mint egyes gnek szintjn szoks, mivel az egyes gnek hossza a legtbbszr nem elgg hossz ahhoz, hogy a Markov-lnc rendjt meg lehessen hatrozni. Így pldul egy 1000*bp* hosszsg szekvencia esetn csupn a msodrendű Markov-jelleget lehetne kimutatni, az viszont valsznűtlen, hogy bzisok s bziskettesek frekvencija kellően lerja a gnt. (Megjegyzem, hogy azta jval hosszabb gnek is ismertek, s ugyanakkor a Markov-lnccal trtnő modellezs elvesztette az rdekessgt.) A modell alkalmas arra, hogy egyes rszletek, bzis-n-esek gyakorisgt megjsoljuk. Pldul, ha kiszmoljuk egy restrikcis szekvencia[4] gyakorisgt a genomban, megmondhatjuk, hny kromoszmadarab keletkezik, amikor egy adott restrikcis enzim műkdsbe lp a genomon. Becslst adhatunk arra is, vajon milyen hossz lesz a leghosszabb ismtlődő szakasz valamely szekvenciban.

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 gondolatmenenetet −, hogy az angol, és általában a természetes nyelvek szintaxisa nem írható le Markov-folyamatokkal, ill. regulris grammatikval, csupn krnyezetfggetlennel (ld. a 3.6.1. fejezetben). A kvetkező kt fejezetben olyan eredmnyeket ismertetek, amelyek a Markov-modellek elgtelensge mellett hoznak fel bizonytkokat a DNS-szekvencik esetben is.

3.2 A Zipf-analzis

G. K. Zipf mr az 1930-as vekben vgzett szgyakorisgi vizsglatokat termszetes nyelven rott szvegekkel (Zipf [1935, 1949]). sszeszmolta nagy korpuszokban az egyes szavak előfordulsi gyakorisgt, s sorrendbe helyezte az egyes szavakat gyakorisguk szerint: $R = 1$ jelentette a leggyakoribb szt, $R = 2$ a msodik leggyakoribbat, s gy tovbb. Ezutn brzolta az ω gyakorisgot (frekvencit) az R sorrend fggvnyben. Az eredmny meglepő volt: az $\omega(R)$ fggvny nem exponencilisan csengett le, ahogy vrnnk, hanem hatvnyfggvnyszerűen:

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

ahol a ζ kitevő, nyelvtől s a korpusz tartalmtl fggetlenl, univerzlisan 1 krli rtket vett fel: $\zeta \approx 1$. Ez azt jelenti, hogy log-log skln brzolva, az $\omega(R)$ fggvny egy $-\zeta$ meredeksgű egyenest ad. A Czirk et al. [1995] ltal idzett irodalom szerint, a Zipf-analzist elvgeztk vrosok, valamint ipari vllalatok mret szerinti eloszlsa vizsglatra is.

[4] Olyan szekvencia, amelyhez egy restrikcis enzim kapcsoldhat, s gy elvghatja a szekvencia kzelben a DNS-szlat.

A mdszer alkalmazsa DNS-szekvencik esetn felveti azt a krdst, hogy vajon mit tekinthetnk "sznak" a genomban? A kdol szekvencik esetn knnyen addik a vlasz: a tripletek (kodonok) a legkisebb "rtelmes" ptőelemek. Viszont a nem kdol szaka-szok jelentős rsznek funkcija mg nem ismert, ezrt Mantegna et al. [1994, 1995b] bevezeti a Zipf-analzis egy sszerű ltalnostst, amelyet magyarul "szimblum n-es Zipf-analzisnek" nevezhetnk (n-*tuple Zipf-analysis*). Ennek lnyege az, hogy egy n szimblum hosszsg "ablakot" mozgatunk vgig a szekvencin, minden lpsben egyetlen (teht nem n) karakterrel elmozdtva az ablakot, s az gy kapott jel n-esek frekvenciira szmtjuk ki az $\omega(R)$ gyakorisg-sorrendben elfoglalt hely fggvnyt.

A szimblum-n-es Zipf-analzis termszetes nyelvi szvegek esetben ppgy hatvnyfggvnyt eredmnyez, mint a hagyomnyos Zipf-analzis, viszont eltrő lesz az exponens. Mantegna et al. [1995b] egy angol nyelvű enciklopdia cikkeit elemeztk ilyen eszkzkkel: a hagyomnyos, szavak gyakorisga alapjn vgzett Zipf-analzis $-0,85$ kitevőt eredmnyezett, mg a betű-n-esek ($n = 3, 4, 5, 6$) $-0,57$ krli meredeksgekkel jellemezhető brkat produkltak a $10 \leq R \leq 1000$ kt nagysgrendet tfog intervallumban. A kt exponens klnbsgt a szerzők abban ltjk, hogy mg a hagyomnyos Zipf-analzisben csupn egy meghatrozott szkincs elemei vesznek rszt, addig az n-esekkel vgzett Zipf-analzis sorn kibővl ez a halmaz.

Kiprbltk ezt a szimblum-n-es Zipf-analzist mestersges nyelvekre, mghozz a UNIX opercis rendszer kompilllt fljaira is ($n = 6, 8, 10, 12$). A 9000000 bit hosszsg szveg egy ktelemű bcből ($\{0; 1\}$) pl fel, s a DNS-adatokkal val knnyű sszevethetősg rdekben (ngy elemű bc) vlasztottk meg gy az n-eket. Ebben az esetben is lineris brkat kaptak a ktszer logaritmikus skln, s a linearitsi tartomny szrevehetően nőtt n nvelsvel. Az $n = 12$ mellett $\zeta = 0,77$ exponenst kaptak a szerzők.

Mantegna et al. [1994] s Mantegna et al. [1995b] az akkoriban hozzfrhető leghosszabb emlős, gerinctelen s virlis DNS-szekvencikra alkalmazta a Zipf-analzist. (Prokariota szervezetek, illetve mitokondriumbl s kloroplasztiszbl szrmaz szekvencikkal nem foglalkoztak.) Ahhoz, hogy egy L hosszsg szakaszt n-es Zipf-analzisnek lehessen alvetni, $L \gg 4^n$-nek teljeslnie kell (a cikkben $L > 10 \cdot 4^n$ teljesl). A cikkben az 50000bp-nl hosszabb, vagy az ezt az rtket megkzeltő hosszsg szekvencikat hasznltak fel, gy az $n = 3, 4, ...7$-eseket lehetett csak vizsglni, de a klnbző n-ekre nem kaptak lnyegesen eltrő eredmnyeket. Nyitott krds, hogy vajon a ζ rtke konstans vagy monoton nő, ha nveljk n rtkt. E krds megvlaszolshoz (ζ sklzsi tulajdonsgainak feldertshez) az 1995-ben hozzfrhető szekvenciknl jval hosszabb szekvencikra lett volna szksg. Fontos mg megjegyezni, hogy az adatbzis felhasznlt ttelei messze nem reprezentatvak az lővilg egszre nzve: a GenBank adatai kztt nagyobb arnyban szerepelnek a kdol szakaszok, mint az lővilgban ltalban, egyes fajok, csoportok adatai tlslyban vannak ms csoportokkal szemben, tovbb gyakran szerepelnek az adatbzisban klnbző

fajok rokon szerepet betltő szekvencii is.

Ezen hossz DNS-szekvencik Zipf-analzise a Zipf-trvnyben megszokott hatvnyfggvnyt eredmnyezett. Pldul, az emberi HUMRETBLAS szekvencia $\zeta = 0,33;\ 0,34;\ 0,34$ exponenst adott rendre $n = 4; 5$ s 6 mellett, az $R \leq 4^{n-1}$ tartomnyban. Az $R \geq 4^{n-1}$ rgiban gyorsan leesik a frekvencia-fggvny, mivel nhny bzis n-es nem fordul elő a 180 000 karakterből ll szekvenciban.

A kvetkező lpsben a szerzők az egyes szekvencikat kt-kt jelsorozatt bontottk szt: a GenBank-beli fjl informcii alapjn klnvlasztottk az egyes szekvencik ismert kdol rgiit a tbbi szakasztl. A kdol s a nem kdol[5] genom-elemek konkatencii (sszefűzsei) erősen eltrő Zipf-viselkedst mutattak. A nem kdol szekvencik a termszetes nyelvi szvegekhez hasonl, (3.2) alak hatvnyfggvnyt eredmnyeztek az $R \leq 4^{n-1}$ intervallumban. Ezzel szemben, a kdol szakaszok Zipf-analzise az $R \leq 4^{n-1}$ tartomnyban *logaritmikus* fggvnnyel kzelthető grbt adott, amely a lin-log skln lineris:

$$\omega = b - c \log_{10} R. \tag{3.3}$$

Ezt az eredmnyt Mantegna et al. [1995b] kellő szm s kellő hosszsg szekvencira kimutatta, $n = 3, 4, ..7$-re. rdemes megjegyezni, hogy e cikk s Kanter & Kessler [1994] szerint (3.3)-hoz hasonl eloszlst kvetnek az bc betűinek gyakorisgai az emberi nyelveken rt szvegekben is.

Mit jelent ez a megfigyels? Czirk et al. [1995, 1996] vizsglja azt a krdst, hogy milyen tpus "szvegek" milyen Zipf-fggvnyt eredmnyeznek. Levezeti, majd szimulcival is igazolja azt, hogy a Markov-modellek Zipf-fggvnyei lpcsőzetesen futnak le, szemben a klnbző mdszerekkel előlltott, hossz tv korrelcit mutat szekvencik hatvnyfggvny szerinti lecsengseivel. Utbbiak jl kzelthetők Markov-modellekkel a kzpső tartomnyban, de a Zipf-plot kt szle lnyegesen eltr az illesztett Markov-modellből szmtott grbtől. (Msik magyarzatot javasol a Zipf-trvny s a Markov-folyamatok sszekapcsolsra Kanter & Kessler [1994].)

Vajon illeszthető-e Markov-modell a DNS-szekvencikbl szmtott $\omega(R)$ fggvnyekre? Mantegna et al. [1995b] szerint az egyes szekvencikbl empirikusan szmthat $p_{\sigma_1\sigma_2}$ tmeneti mtrix s v_σ gyakorisgi vektor elemei alapjn

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

[5] A "nem kdol szakasz" kifejezs itt, az eddigiekben s az elkvetkezőkben is, sokszor azt jelenti, hogy az adott szakasz valdi szerept nem ismerjk, vagyis nincs kizrva az sem, hogy egyes rszei proteint kdoljanak. De valsznűleg ms a szerepe, ha van egyltaln.

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

A tovbbi ksrletek azt mutatjk, hogy a Markov-lnc rendjnek nvelsvel 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-analzis 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 mennyisg vltozsa statisztikailag "sszefgg": ha az X adatsor i-ik eleme, X_i nagyobb, mint az X-ek $\langle X\rangle$ tlaga, akkor "ltalban" Y_i is nagyobb, mint az $\langle Y\rangle$ tlag. Egzaktabb formban:

$$\langle X \cdot Y\rangle - \langle X\rangle\langle Y\rangle > 0. \tag{3.5}$$

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

$$\langle X \cdot Y\rangle - \langle X\rangle\langle Y\rangle < 0. \tag{3.6}$$

Egy "hossz" X_i szimblumsorozat esetn beszlhetnk *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} vltozsa korrellhat, klnbző i-kre. A

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

fggvnyt, ahol az tlagolsokat az i pozcikra 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 beszlhetnk,

ha ltezik olyan R *karakterisztikus tvolsg*, amely szerint $C(k)$ exponencilisan lecseng (pldul Markov-folyamatok esetn):

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

Amennyiben nem ltezik ilyen R karaktertisztikus tvolsg, hanem $C(k)$ hatvnyfggvny szerint cseng le,

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

akkor *hossz tv korrelci* van jelen a szimblumszekvenciban. Kimutathat, hogy n-ed rendű Markov-folyamatok nem produklhatnak hossz tv korrelcikat, azok lecsengenek n nagysgrendjben.

Hossz tv korrelcik ltt tbb mdon is kimutathatjuk valamely szekvenciban. Az első megolds a $C(k)$ fggvny felvtele kzvetlenl, az adatokbl. Ennek viszont meglehetősen nagy a komputcis ignye. Viszont az albbiakban bemutatand "vletlen bolyongs-mdszer" egy olyan mennyisget vezet be, amely szintn kimutatja a hossz tv korrelcikat, viszont diszkrt Fourier-analzis s egy "cssztatott doboz" (*sliding box*) trkks felhasznlsval knnyen szmthat.

A *vletlen bolyongs-modellt* először 1992-ben C.-K. Peng s trsai, tbbsgk a Bostoni Egyetem munkatrsai (Peng et al. [1992], Stanley et al. [1992, 1993a,b], Peng et al. [1993]) alkalmaztk DNS-szekvencikra. Később ms fizikusok felhasznltk a mdszert írott szvegek elemzsre is (Schenkel et al. [1993], Amit et al. [1994], Dietler & Zhang [1994], Ebeling & Neiman [1995]).

Az eljrs lnyege a kvetkező: kpzeljnk el egy bolht, amelyet egy szmegyenes origjba helyeznk. Felolvassuk neki a szekvencit, s amennyiben pirimidinvzas bzist hall (`C` vagy `T`), gy előre ugrik egyet, míg purinvzas bzis esetn (`A` vagy `G`) htra ugrik. Ms szablyok is elkpzelhetők, pldul `A` esetn (ill. `C` esetn, `G` esetn, `T` esetn) lp csak előre a bolha, mg az sszes tbbi bzis esetn htra; vagy erős hidrognkts esetn (`C` vagy `G`) lpnk előre, gyenge hidrognkts esetn pedig htra; vagy az n. hibrid (vagy alfabetikus) szably esetn: `A` s `C` jelenti az előreugrst, `G` s `T` pedig a htraugrst. De a "purin-pirimidin"-szably alkalmazsa a legelterjedtebb, a kvetkezőkben n is erre fogok hivatkozni. Elkpzelhető persze az is, hogy valaki az egyes bzisokhoz nem egsz nagysg ugrst rendel (pl. a molekulris tmeget vagy a hidrofobicitst javasolja a szakirodalom).

Amennyiben u_i-vel jelljk az i-ik lpst ($u_i = \pm 1$), gy a bolha helyzete az i-ik lps utn nyilvn

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

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

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

ahol

$$\Delta y(l) = y(l_0 + l) - y(l_0), \tag{3.12}$$

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 fggvnnyel:

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

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

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

s az eltrs az α-exponensben nyilvnul meg:

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

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

• Hossz tv korrelcik esetn viszont, $C(l)$ nem jellemezhető R karakterisztikus hosszal, (3.9) szerint viselkedik, s ezrt $F(l)$ α-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 eljrs tvihető termszetes nyelven írott szvegekre is. Ebben az esetben egy-az-egyhez kdolst alkalmaztak, azaz tírtk a szveget egy binris bc segítsgvel (szemben a "DNA-walk"-kal, ahol is kt-kt bzist a legtbbszr ugyangy kdoltak). Pldul t biten kitűnően lehet kdolni az angol bc 26 betűjt, a szkzt, s a legfontosabb írsjeleket.

Az eredmnyek nagyon rdekesek. A vizsglt DNS-szekvencik nagyon jl sztvlaszthatk kt csoportba (Peng et al. [1992]): a tisztn kdol szakaszokbl, exonokbl ll szekvencik $\alpha = 0,50 \pm 0,01$ exponenssel jellemezhetők, azaz nem mutatnak hossz tv korrelcikat. (Ezek a szekvencik egyrszt olyan, alacsony rendű fajokbl szrmaz gnek szekvencii, amelyek nem tartalmaznak intronokat; msrszt cDNS-szekvencik, azaz (ld. pl. Weaver & Hedricknl) olyan dezoxi-ribonukleinsav-szekvencik, amelyeket mRNS-ekből, reverz transzkripcival lltottak elő, abban a fzisban, amikor az elsődleges tiratbl (*primary transcript*-ből) az intronokat mr kivgtk a megfelelő sejtszintű folyamatok.) Viszont az intront is tartalmaz szekvencik esetben az $F(l)$ fggvny exponense $\alpha = 0,61 \pm 0,03$. (Peng et al. [1992]) A kt rtk erősen szignifikns eltrse mig nincs kellő alapossggal megmagyarzva.

Írott szvegek esetben, a vizsglathoz hasznlt szvegek kztt megtalljuk a Biblia eredeti, hber nyelvű szvegt, valamint modern fordítsait, tovbb Shakespeare-drmkat, regnyeket, sőt egy sztrat is. Nhny rdekesebb eredmny:

• A szvegek sok nagysgrenden keresztl lland α-exponenssel jellemezhetőek, melynek rtke ltalban $0,6 - 0,7$ kz esik (Schenkel et al. [1993]).

• A kitevő rtke nem jellemző a szerzőre, hiszen pldul a *Hamlet* exponense $0,56$, míg a *Rme s Jli* $0,6$ (Schenkel et al. [1993]).

• A fordítsok sorn, gy tűnik, a korrelci mrtke cskken. Habr a Biblia exponense kimagaslan magas ($\sim 0,75$), a fordítsai sorn ezen rtk szisztematikusan cskken.[6] (Amit et al. [1994])

• A sztr a ltszlag fggetlen szcikkek hosszn jval hosszabb korrelcikat mutat, amely jelensg nem magyarzhat pusztn a tartalom korrelltsgval. Ezek szerint a sztr szcikkeinek elrendezse "nem vletlenszerű", előzetes sejtsnkkel ellenttben, írja (Schenkel et al. [1993]).

• A szvegeket kisebb rszletekre, pldul szavakra, mondatokra vagy $l = 10 - 10000$ karakter hosszsg, azonos darabokra szabdalva s ezeket a darabkkat szabadon permutlva, a szabdals hossznak nagysgrendjig megmaradnak a korrelcik, azon tl pedig eltűnnek (Schenkel et al. [1993]).

Lefordított szmítgpprogramokat (`.exe`-file-okat) is megvizsgltak, ezek esetben $0,9$-et is meghalad kitevőt talltak (Schenkel et al. [1993]). rdekes mg a "humn vletlenszmgenertor"

[6] Nem talltam a szakirodalomban vilgos vlaszt arra a krdsre, hogy a magas hatvnykitevő mirt nem lehet a hber nyelvnek vagy ortogrfinak a jellegzetes tulajdonsga.

vizsglata: az egyik szerző 0 s 9 kztt írt le szmokat, "vletlenszerűen". A vletlenszerű eloszlsra 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 hossztv 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 esetn az $S(f)$ teljestmnyspektrum szintn hatvnyfggvny:

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

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

$$\alpha = \frac{\beta + 1}{2}. \tag{3.16}$$

A DNS-szekvencik diszkrt Fourier-analzisből kapott eredmnyek jl visszaadjk ezt az elmleti sszefggst is.

Miknt hasznlhatk a hossz tv korrelcik a kdol s a nem kdol szakaszok klnvlasztsra? Stanley et al. [1993a] nyomn tekintsnk egy m hosszsg "megfigyelsi dobozt" (az emltett cikkben $m = 800bp$), s mozgassuk a dobozt a DNS-szekvencia mentn. Minden lpsben kiszmoljuk a dobozon belli szekvencinak a fentiekben bevezetett α exponenst, s brzoljuk ezt az rtket a doboz pozcijnak a fggvnyben. Az idzett cikk 9. brja illusztrlja ezen j algoritmus ltal produklt fggvnyt az lesztő III. kromoszmjnak feldolgozsakor: 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 korrelcikat a kdol szakaszokban lvő klnbző kodonok nem egyenletes eloszlsra vezeti vissza (Herzel et al. [1997a], Herzel & Große [1995, 1997b], Große et al. 1997]). Bevezetnek egy "klcsns informci-fggvnyt" (*mutual information*), amelynek segtsgvel jabb algoritmust javasolnak a kdol s a nem kdol szakaszok klnvlasztsra.

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

Az informcielmlet Shannon-fle megkzelítsnek a clja az, hogy elvi hatrt adjon arra, hogy milyen sebessggel lehet tovbbtani valamely "nyelv" "mondatait" egy (zajnlkli vagy zajos, vges sok llapottal rendelkező vagy folytonos) csatornn keresztl. Ebben a megkzelítsben egy "mondat" hírrtke a valszínűsgtől fgg: minl valszínűtlenebb a mondat, annl nagyobb a hírrtke. (Gondoljunk csak az jsgírtanfolyamok alapttelre: az nem hír, ha a kutya megharapta a postst, de az igen, ha a posts megharapta a kutyt.) A biztos esemnynek nincs hírrtke, mert azt nem is kell tovbbítani, elegendő, ha a vevő oldalon automatikusan hozztesszk a vett zenethez. A negatív logaritmussal trtnő definícit pedig az motivlja, hogy a kt kapcsol segítsgvel, kt lyukkrtyn vagy kt floppyn tovbbíthat informci mennyisge ktszerese legyen az egy esetn tovbbíthatnak.

Shanon, kvetkező lpsben bevezeti a forrs H entrpijt:

$$H = -\sum_i p_i \log(p_i), \tag{3.17}$$

ahol p_i az i-ik esemny bekvetkeztnek, az i-ik karakter előfordulsnak a valsznűsge. Az entrpia annl nagyobb, minl tbb esemny, minl homognebb valsznűsgekkel oszlik el. Ez az oka annak, mirt olvashatunk ritkn az jsgban kutyaharapsrl, de annl gyakrabban a lotthzs eredmnyről: az előbbi esetben kt, egy valsznűtlen s egy nagyon valsznű esemnyről van sz, mg az utbbiban sok, egyenlően valsznű esemnyről.

Amennyiben a forrs nem egyszerűen fggetlen esemnyeket llít elő, hanem egy korrellt szimblumsorozatot (mondatokat) produkl, akkor a forrs entrpija:

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

ahol $P(Q)$ az n betűből ll Q sztring előfordulsi gyakorisga. (Σ^n jelenti a Σ elemeiből alkotott, n hosszsg lncok halamzt.)

Az informcielmlet tovbbi elemeire nem lesz szksgnk. Az "entrpia" elnevezs a fizikbl szrmazik. Az analg kpleten tl, tovbbi kapcsolat is kimutathat: egy gz fizikai entrpija (definícis konstans szorztl eltekintve) megegyezik a rszecski tr- s sebessgkoordintinak informcielmleti entrpijval.

Ezen kitrő teszi rthetőv Herzel s Große [1995] *klcsns informci-fggvnyt* (*mutual information*). Jellje a λ elemből ll $\{A_1, A_2, ..., A_\lambda\}$ bcnk esetn $p_{ij}(k)$ annak egyttes valsznűsgt, hogy valamely pozciban az A_i szimblumot talljuk, k pozcival később pedig az A_j betűt. Jellje p_i annak a valsznűsgt, hogy egy nknyes n pozciban az A_i szimblum fordul elő, mg q_j annak a valsznűsgt, hogy az $n+k$-ik helyen A_j tallhat. Mivel a szekvencinkrl feltesszk, hogy stacionrius, $p_i = q_i$ minden $i = 1, 2, ...\lambda$-ra. Kt szimblumot akkor s csak akkor neveznk *statisztikailag fggetlennek*, mint arrl mr volt sz, ha $p_{ij}(k) = p_i \cdot p_j$. Az

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

mennyisg, amelyet *klcsns informcinak neveznk*, a statisztikai fggetlensget, ill. a korrelcik erőssgt mri.

Herzel s Große kutatsai szerint klnbző lőlnyek kdol DNS-szekvenciiban megfigyelhető egy nagyon jellegzetes 3-hosszsg oszcillci a lassan lecsengő $I(k)$ fggvnyben, mg nagy tvolsgokon ($k > 1000bp$) is: ha k hrommal oszthat, $I(k)$ tbbszrse a szomszd k-k esetn felvett $I(k)$ rtkeknek. Ennek okt a szerzők abban lelik meg, hogy eltrő az egyes bzisok gyakorisga a kodon (triplett) első, msodik, ill. harmadik pozcijban. Nem kdol szekvencik esetn ez a periodicits nem figyelhető meg.

Große et al. [1997] bevezetik az *AMI tlagos klcsns informci* (*Average Mutual Information*) nevű mennyisget, amely az $I(k)$ fggvny rtknek tlaga egy N hosszsg (pl. a $k \in [5; 5 + N]$) intervallumban ($N \approx 50$-től működik az eljrs). Nem kdol szekvencikra az AMI lecseng N^{-2}-vel, mg a kdol szekvencikra az AMI egy pozitv rtkhez tart, amely rtk a bzis-gyakorisgok kodon-pozcitl val fgggsből szrmaztathat.

Ezen megfigyels egy olyan algoritmushoz vezet, amely a kdol s a nem kdol szekvencikat a tbbi algoritmushoz hasonl megbzhatsggal kpes sztvlasztani, nem ignyel előzetes tantst (tbb algoritmus is ltezik, amelyek a neuronhlzatoknl megszokott előzetes "training"-ből szrmaz paramtereket hasznlnak fel), s elg j "finomsggal" ($\approx 50bp$ hosszsg doboz pontossgval) működik.

3.5 Redundancia

Az emberi nyelvek kzismert tulajdonsga a redundancia. Ez teszi lehetőv, hogy a kommunikcit ne zavarjk olyan "zajok", mint a beszlő beszdhibja, nyelvbotlsa, rott szveg esetn az elgpels, s gy tovbb. Ez a redundancia jellemző a genetikai kdra is, mint ahogy azt mr megemltettem.

Shanon *relatv entrpinak* nevezi egy forrs (3.17)-ben definilt entrpijnak, s az azonos szimblumok segtsgvel elrhető maximlis entrpinak az arnyt (Shanon & Weaver [1986]). Ez a mennyisg a maximlis lehetsges kompresszit jelenti, ha ugyan azon bcben kdolunk. A *redundancia* a relatv entrpit egszti ki 1-re. Mantegna [1995b] jellsei szerint az *n-gramok entrpija* (λ elemű bcre)

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

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

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

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

Nagy Ferenc [1986] becslsknt 68%-ot szmol egy tlagos nyelv redundancijra (32 betű, 10 betűs szavak esetn). Shanon a htkznapi angol nyelv redundancijt durvn 50%-nak becsli. Ezt az eredmnyt adja mind a nyelv Markov-lncokkal trtnő approximcijn alapul entrpiaszmts, mind a titkosrsok elmletnek nhny eredmnye. Ezt tmasztja al az is, hogy a ksrletek szerint egy angol szveg ltalban rekonstrulhat, ha a karakterek felt vletlenszerűen trljk. Emltsre mlt mg Shanon eszmefuttatsa arrl, hogy milyen mrtkű redundancia esetn lehetsges keresztrejvnyt kszteni: ha tl sok megszorts vonatkozik a nyelvre, nyilvn nehz nagy keresztrejtvnyt szerkeszteni. Viszont zrus redundancia esetn brmely betűsorozat ltező szt ad, vagyis a karakterek tetszőleges kt dimenzis elrendezse keresztrejtvny. Shanon szerint 50%-os redundanciig lehetsges kt dimenzis, mg 33%-os redundanciig hrom dimenzis rejtvnyt kszteni, ha a nyelvre vonatkoz megszortsok vletlen termszetűek.

Mantegna et al. [1994, 1995b] klnfle szekvencikra, valamint ezen szekvencik kdol, ill. nem kdol szakaszainak konkatenciira szmtotta ki a (3.20) szerinti $H(n)$ entrpit. Mivel rtkelhető eredmnyt csak $L > 100 \times 4^n$ esetn kaphattak, $n \geq 8$ rtkeket mr nem vizsglhattak, sőt egyes szekvencik esetn mg kisebb n rtkeknl meg kellett llniuk. A (3.21)-beli konvergencia nagyon lass, s ilyen kis n-ek esetn a hosszabb repeat-szekvencik hatst sem vizsglhatjuk. Ennek ellenre emltsre mlt eredmnyekre jutottak: nhny kivteltől eltekintve,[7] a nem kdol szekvencik redundancija szignifiknsan magasabb a kdol szekvencik redundancijnl, mg a komplett genomdarab grbje e kettő kztt halad az egyikhez kzelebb attl fggően, hogy a kdol vagy a nem kdol elemek vannak-e tbbsgben az egsz genomban.

Mantegna et al. e megfigyelsből is azt a kvetkeztetst vonjk le, hogy a nem kdol szekvencik, szemben a kdol elemekkel, a statisztikai tulajdonsgaik tekintetben a termszetes nyelveken rott szvegekre hasonltanak.[8]

3.6 Modellezs generatv nyelvtanok segtsgvel

[7] A kivtelknt kzlt brbl kiderl, hogy ebben az esetben csak $n \leq 4$-re lehetett a $H(n)$-eket szmtani, s a grbk "trendje" alapjn asszimpttikusan itt is lehetsges, hogy a fenti megfigyels megllja a helyt.

[8] "Linguistic features of noncoding DNA sequences", mint azt az egyik cikkk cmben is emltik.

Az eddigiekben a DNS-szekvencik sok olyan statisztikai tulajdonsgval tallkoztunk, amelyek a termszetes (emberi), valamint a szmtgpes nyelvekkel rokontjk a DNS-szekvencikat. Arra is lttunk tbb rvet, hogy a szimblumszekvencik hagyomnyos sztochasztikus modelljei, a Markov-folyamatok nem rjk le kielgtően az emltett megfigyelseket. Ezrt az albbiakban egy msik modellt fogok javasolni, a generatv grammatikk sztochasztikus ltalnostsait, valamint a Zipf-fggvny elmleti szmtst is megmutatom, a modell paramtereiből. A generatv grammatikk elmlete s az automataelmlet nem sztochasztikus fejezetei egybknt az algebra bevett gnak szmtanak, gy itt csupn a szksges fogalmakra fogom korltozni az ismertetst, amelyet az indokol, hogy e tma nem szerepel a fizikus szak tananyagban.

3.6.1 A formlis nyelvek elmletnek 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 Σ halmaz, az ún. *ábécé* fölött, definíció szerint nem más, mint a Σ elemeiből képzett, véges hosszúságú sztringek egy halmaza. (Az n hosszúságú sztring egy rendezett n-est jelent matematikailag.) Jelölje Σ^+ a Σ elemeiből képezett, pozitív (véges) hosszúságú sztringek halmazát, Σ^* pedig ennek kibővítését az e ("empty") üres, azaz nulla hosszúságú sztringgel. Ekkor egy L nyelv Σ fölött nem más, mint Σ^* 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

elemeire fognak utalni. A szóhasználat a konkrét implementációtól függ: szintaxis (mondattan) esetén például Σ 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 Σ^* a konkatenáció műveletével együtt egységelemes félcsoportot 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 $P \to Q$ 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 $(P \to Q) \in F$.

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 $X \Rightarrow^* Y$ csakkor áll fenn, ha X-ből kiindulva, F-beli újraírøszabályok 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:

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

Azt mondjuk, hogy az L nyelvet a G grammatika *generálja*, ha $L = L(G)$. A terminálisok V_T halmazának szerepét a Σ 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 $Q_1XQ_2 \to Q_1PQ_2$ alakú, ahol $Q_1, Q_2, P \in V^*$, $X \in V_N$ és $P \neq e$, kivéve esetleg az $S \to e$ 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 $X \to P$ alakú, ahol $X \in V_N$, és $P \in V^*$.

$i = 3$: Az F minden eleme $X \to PY$ vagy $X \to P$ 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 alapgondolatát 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 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 llapot automatk, illetve az azokkal ekvivalens "llapotcimkzett automatk" (*state labeled automata*, Bird & Ellison [1994]) sztochasztikus ltalnostsai. Teht amikor eddig Markov-lncokkal kzeltettk a szimblumszekvencikat, regulris modelleket 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 "beágyazást" 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] amellett 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. Ezektől 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 mondatszimbólum, 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 mondatszimbólumból 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 → NP VP. Maga az igei csoport is felépülhet igéből (V), tárgyból és határozókból (utóbbiak szintén NP-k, azaz főnévi csoportok): VP → VP NP, illetve: VP → V. A főnévi csoport pedig állhat főnevekből (N) és melléknevekből (A): NP → A NP, és NP → 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 levezetés(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":

$$(...[...)...]$$

.

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 regulárissá 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 regulárissá, 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 létére nyelvészeti magyarázatot találjunk. Ehhez az szükséges, hogy a formális nyelvek elméletét kiegészítsük sztochasztikus eszközökkel. A krnyezetfggetlen grammatikk ilyen ltalnostsait *sztochasztikus környezetfüggetlen grammatikák* (*SCFGs*) néven ismerik. A regulris nyelvek osztlynak sztochasztikus ltalnostsai lnyegben a Markov-modellek, tgabb nyelvosztly ltalnosítsairl pedig nics tudomsom. Utbbiakra viszont egyelőre nincs is szksg, mivel a formlis nyelvek elmletnek legfontosabb felhasznli (nyelvszet, szmtstudomny) leggyakrabban az $\mathcal{L}_2$ nyelvosztllyal 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 mondatszimbólumból az adott szekvenciát. Ezt a valószínűséget úgy értem, hogy amennyiben összefűzzük a "végtelen" sok mondatát, azaz képezünk egy *ideális szövegkorpuszt*, akkor az adott mondat milyen gyakorisággal fordul elő a korpuszban. Az ideális korpusz jó közelítését adhatja természetes nyelvek esetén egy könyv (pl. egy regény), a "DNS-nyelv" esetén pedig a DNS-szekvenciákat tartalmazó adatbázisokat fogom ilyen korpusznak tekinteni.

A nyelvszet hagyomnyos szemlletmdja kt ponton is eltr ettől a megkzeltstől. Előszr is, a nyelvi adatok helyes voltt binrisan kezeli: egy alak vagy grammatikus vagy agrammatikus, vagy eleme a nyelvnek, vagy nem. Ezen megkzelts htrnyait ecseteli meglehetősen irnikus hangnemben, s a statisztikus-sztochasztikus szemllet trnyersrt tőr lndzst Abney [1996]. A msik eltrs pedig abban van, hogy a fonolgiai grammatikalitst nem *korpuszra* (*corpus-based*), hanem a szkincsre, azaz a *lexikonra* (*lexicon-based*) vizsgljk, vagyis nem tesznek klnbsget akztt, hogy pldul egy rendhagy ("tiltott") hangkapcsolat egy gyakori vagy egy ritka szban fordul-e elő. (Ennek az az oka, hogy a hagyomnyos generatv nyelvszet mind a ritka, mind a gyakori szt a nyelv egyenrang elemnek tekinti, s az előfordulsban mutatkoz klnbsgeket a performancihoz, azaz a nyelvszet vizsglati krn kvlre sorolja.)

A krnyezetfggetlen grammatikk sztochasztikus ltalnostst a kvetkezőkppen definil-

hatjuk Krenn & Samuelsson alapjn:

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 mondatszimbólum,*

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

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

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

A $P(X \to Q)$ 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 mondatszimbólumból ("poli-S láncból") levezetett sztring, úgy látható a kapcsolat a korpuszból számított empírikus gyakoriság és a fentebb bevezetett elméleti valószínűség között.

3.6.3 Becsls a Zipf-fggvnyre

Ahhoz, hogy a modell paramtereiből eljuthassunk a Zipf-trvnyben szereplő $\omega(R)$ fggvnyhez, az albbi gondolatmenetet javasolom: szmtsuk ki az egyes terminálisok előfordulási valószínűségét egy "korpuszban", azaz mondatok vgtelen láncában. (Ezt a szmtst tudtommal mg nem vgezte el senki, mivel a sztochasztikus krnyezetfggetlen grammatikkat jelenleg ms clokra, pldul valsznűsgeket is felhasznl parser-ek ksztsre hasznljk.)

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

[9] 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 σ mondatban. (Megjegyzem, hogy habár formálisan nem látom be, de ezen végtelen szummának konvergensnek 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, σ-t eredményező levezetési fák valószínűségeinek[10] az összegét. és jelölje $v_A(a)$ annak a várható értékét, hogy az A-ból sztringekben hányszor szerepel az a terminális:

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

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

$$A \to BC$$

vagy

$$A \to a$$

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 μ mátrixot:

[10] Az ilyen fák valószínűségét szintén az alkalmazott levezetési szabályok valószínűségeinek a szorzataként számíthatjuk ki, akárcsak az S gyökerű fák esetén.

$$\pi_A(a) := P(A \to a)$$
$$\mu_{AB} := \sum_{C \in V_N} \Big[P(A \to BC) + P(A \to CB) \Big]$$

Amikor az A-ból akarom levezetni a σ-t egy CNF alakú grammatikában, két lehetőségem van az elindulásra: amennyiben a σ egyetlen a terminálisból áll, úgy az $A \to a$ szabályt kell alkalmaznom, egyébként pedig az A-t egy $A \to BC$ szabállyal írom újra, és B-ből levezetem σ_1-et, C-ből σ_2-t, ahol $\sigma = \sigma_1\sigma_2$ alakú (e kettő konkatenációja). Ennek megfelelően:

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

(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:

$$\left(\frac{b}{a}\right) = \delta_{b,a}$$
$$\sum_{\sigma \in V_T^*} p_C(\sigma) = 1$$
$$\left(\frac{\sigma}{a}\right) = \left(\frac{\sigma_1}{a}\right) + \left(\frac{\sigma_2}{a}\right)$$

Az eredmény a következő lesz:

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

amelyből:

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

vagyis:

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

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 μ elemeire nincsen semmiféle megkötés ($\pi(\mathbf{a})$ kis megváltoztatására μ is megváltozik), vagyis kis perturbációval megszüntethető egy esetleges szingularitás.

A hagyomnyos Zipf-trvny ppen a szintaxis, mint generatv grammatika, terminlisainak, azaz a szavaknak a most kiszmtott gyakorisgait hasznlja fel.[11] A karakter n-esekre vgzett Zipf-analzisben szereplő frekvencikra becslseket kaphatunk, ha elg hossz szavaink vannak, vagyis a szhatron tnyl karakter-n-esektől első kzeltsben eltekinthetnk.

A következő fejezetben bemutatandó "vektortér-technika" − ahogy M. Damashek [1995], az alapötlet kidolgozója nevezte −, a karakter$n$-eseknek a Zipf-analzisben is szereplő, most megbecslt frekvenciit hasznlja fel.

[11] A hagyomnyos Chomskynus szintaxisban szerepelnek n. *transzformcik* is, amelyek egy generatv grammatika ltal generlt nyelv szavait permutljk, de ezek a permutcik nem vltoztatjk az ltalunk vizsglt gyakorisgokat.

4. fejezet

A vektortr-technika

Az eddigiekben sok olyan statisztikai eszkzkkel megfoghat jelensggel tallkoztunk, amelyekben a kdol s a nem kdol (pontosabban: főleg ismeretlen szerepű, valsznűleg nem kdol) genom-rszletek eltrnek. Feltűnő volt a hasonlsg a nem kdol DNS-szekvencik s a termszetes nyelven rott szvegek kztt. Foglaljuk ssze ezeket az eredmnyeket.

A nem kdol szekvencik s az rott szvegek hatvnyfggvnyknt lecsengő Zipf-fggvnyt eredmnyeztek. Ezzel szemben a kdol szakaszok Zipf-analzisekor exponencilist kaptunk, amelyek az egyes betűk eloszlsra emlkeztettek rott szvegekben. Hossz tv korrelcikat, az rott szvegeken kvl, az intronokat tartalmaz gnekben is talltunk. Az informcielmleti redundancia szintn e kt tpus szimblumszekvenciban magas, szemben a "lehetősgeit" jl kihasznl exonokkal.

Az ilyen klnbsgekre plő, exonokat kereső algoritmusokat a "mintzat-kereső" mdszerek kz sorolnm (ld. a 2.2 fejezetben), mivel egy előre definilt tulajdonsgot (hossz tv korrelcik ltt, hatvnyfggvnyt kvető Zipf-plotot, magas redundancia-rtkeket) keres. Az albbiakban bemutatsra kerlő mdszer ezzel szemben a "visszakereső algoritmusok" kz tartozik, amennyiben a szekvencik kztti hasonlsgot mri. Az eljrst eddig csak rott szvegekre alkalmaztk, DNS-szekvencikra mg nem, gy nem trivilis az, hogy egy ilyen hasonlsgi mrtk jelezni kpes a funkcionlis (kdol vs. nem kdol) klnbsgeket. Mgis, mint az ki fog derlni az eredmnyeimből, a javasolt mdszer kivlan klnvlasztotta az exonokat s az intronokat. A mdszer sikere vlemnyem szerint sszefgg Clay et al. [1996] s a tbbi, a 3. fejezet elejn emltett cikk eredmnyeivel, amelyek szerint habr van korrelci az azonos eredetű exonok s intronok bzisfrekvencii kztt, de azok szisztematikusan eltrnek egymstl.

4.1 A Damashek-mdszer 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 *Acquaintance*nek 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 kapacits vges 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.[1] A szövegből képezett $\mathbf{x}$ vektor i-ik komponensét éppen az m_i lenormálásából kapott frekvenciaként képezzük:

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

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

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

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 $d(\mathbf{x}, \mathbf{y})$ pozitv, 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.[2]

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

[1] Fickett [1996] is emlt hasonl mrtket, de ő az n karakter hosszsg ablakot minden lpsben n karakterrel mozdtja el, vagyis a *nem* tfedő karakter-n-esek frekvenciit hasznlja.

[2] Legyenek pldul $\mathbf{x}$, $\mathbf{y}$ s $\mathbf{z}$ egy skban fekvő vektorok ilyen sorrendben. A szomszdos vektorok zrjanak be 30°-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. tblzat: Angol nyelvű fizikai (Ph), egyb tmj angol szvegek (E), francia (Fr) s magyar (H) levelek (`.txt`-fjlok) alapjn ksztett, a karakterhrmasok gyakorisgt jellemző ($n = 3$) vektorok szorzata. Lthat, hogy az azonos nyelvű szvegek egyms kztt szignifiknsan magasabb szorzatokat adnak, mint a klnbző nyelvű szvegek szorzatai. A mdszer tma szerinti vlogatsra is alkalmas, hiszen kt E-szveg, ill. kt Ph-szveg hasonlsga 15%-kal magasabb, mint egy E- s egy Ph-szveg szorzata.

magam például öt, fizikai témáj 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 ($r = 29$-fle 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 hasonlsgi 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 λ slyt az egyes vektorokhoz, kisebbet a rvidebbekhez s nagyobbat a hosszabbakhoz, majd a kt dokumentum ezen slyval, mint szorzfaktorral, korrigljuk a centroid vektorok levonst kvetően kapott szorzatot.

4.2 A Damashek-mdszer tovbbfejlesztse s diszkusszija rott szvegekre

A mdszert inkbb ms irnyba fejlesztettem tovbb: összefűztem mind a négy szövegtípusból (E, Ph, Fr, H) néhány szöveget egyetlen fűzérré.[3] 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 eljrsra "mozg dobozos mdszer"-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 mdszer alkalmasnak tűnik arra, hogy klnbző jellegű szvegekből sszetett szekvencikat a komponenseire bontsuk szt, legalbbis olyan pontossggal, amilyet a mozg doboz mrete lehetőv tesz.

Az eredményt rviden azzal magyarzhatjuk, 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-mdszer sikert hrom tnyezőre vezethetjk vissza: kt nyelvi s egy nyelven kvli tnyezőre.

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

A msodik tnyező a nyelv fonolgija, mghozz a hangtan *fonotaktiknak* nevezett ga, amely a markovi tulajdonsgokat, vagyis a lehetsges hangkapcsolatokat rja le. A kutatsok ezen irnya előrbb tart, mint az előző bekezdsben emltett krdsek, de mivel nem kvnok nyelvszeti problmkba 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

[3] 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 kezdőpozcijnak 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 *tere***mt** szban. Ezrt ezeket is agrammatikusnak tekintjk. Egy *corpus-based* statisztikus megkzelts klnbsget tud tenni a gyakori *teremt*, *bara***ck**, *rece***pt** szavak kevsb agrammatikus

volta, a ritka s idegen hangzs **ng***anaszn*,[4] **scs**, **fjl** 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-modellel,[5] amely N állapotot $(s_1, s_2, ..., s_N)$ és M darab jelet $(\sigma_1, \sigma_2, ..., \sigma_M)$ 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 ergodikus a Markov-modellünk; 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 $\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}$ szimbólum n-es elméletileg jósolt valószínűsége, azaz a Damashek-vektor megfelelő komponense:

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

Mivel a p_{ij} és a_{ij} paramétereket 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 $P(\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n})$-k empírikus közelítései, valamint felhasználhatjuk a jelek empírikus gyakoriságait. Az i-ik jel ν_i frekvenciájára adható elméleti becslés:

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

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

---

[4] Egy szamojd np s nyelv neve.

[5] Felmerlhet, hogy miért nem egyszerűsítem le a gondolatmenetemet Markov-láncokra. Két okból ragaszkodtam a Markov-modellhez. Egyrészt, a reguláris grammatikákat − amelyekkel lerhatjuk a fonolgit Kaplan & Kay [1994] szerint − Markov-modellekre, 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.

Miutn diszkutltuk a Damashek-mdszer sikernek kt nyelvszeti tnyezőjt, emltsk meg a harmadik, nyelvszeten kvli tnyezőt is, a helyesrsi hagyomnyt. 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 szvegekben '*s*'-knt, az angolban leggyakrabban '*sh*'-knt, a franciban '*ch*'-knt, mg a nmetben 'sch'-knt jelenik meg. Eme tnyező szerept sem szabad figyelmen kvl hagyni, hiszen a nmet szvegekből kszlt vektorok meghatroz komponensei lesznek az '*sch*'-t tartalmaz komponensek, amelyek rtke a magyarban szinte zrus lesz. A helyesrsi hagyomny hatst fonetikai-fonolgiai trs alkalmazsval lehetne kikszblni, de egyelőre nem llnak rendelkezsemre kellő szmban elg hossz ilyen tpus szvegek.

4.3 A szmtsi algoritmus

A szksges szmtsokat C-nyelvben, UNIX opercis rendszer al rt programokkal vgeztem el.

Először a feldolgozand szimblumszekvencikat tartalmaz fjlokat ksztettem el, e-mail levelek mint `.txt`-fjlok felhasznlsval, ill. a *GenBank*-ből letlttt genomrszletek megfelelő sztakaszainak a megkeressvel. Az első program, amelyet rtam (`vektor`), ezekből a fjlokbl – a paramterknt megadott n-re – elksztette s fjlba mentette az n-gram-frekvencikat tartalmaz vektorokat. Írott szvegek esetn a tbb, egymst kvető res karaktereket (*space*-eket) ki kellett hagyni, hiszen ezek a nyelvileg irrelevns ismtlődő szekvencik indokolatlanul dominlhatnk a keletkező vektor jellegt, ill. a skalrszorzatokban kiugran magas, de a szvegek kategorizlsa szempontjbl nem indokolt tagokat eredmnyezhetnnek.

A vektorok elksztse sorn felhasznltam Damashek [1995] ltal javasolt trkkt. Mikzben az n karakter hosszsg ablakommal vgighaladtam a szekvencin, nem volt rdemes egy r^n mretű vektorba gyűjttteni az egyes n-esek előfordulsnak a szmt (ahol r a szekvenciban előfordul karakterek szma, az "bc" mrete), hiszen főleg az $L \approx 1000 - 5000$ hosszsg rott szvegekre ($r = 29$) $L \ll r^n$. Damashek javaslatt kvetve, az egyes n-eseket talaktottam egy-egy szmm (egy r alap szmrendszerben felrt szmnak tekintve a karakter-n-est), majd ezt az jabb szekvencit rendeztem a "gyorsrendezs" (*quicksort*) nevű algoritmus segtsgvel (Press et al. [1989]).[6] Ezt kvetően mr knnyű sszeszmolni s fjlba menteni a nulltl eltrő gyakorisgokat (vektorelemeket).

Az gy kapott fjlokat hasznlta fel a `szorzas` nevű programom, amely a vektorok skalris szorzst vgezte el, s az eredmnyt a standard outputon jelentette meg. Írtam egy segdpro-

[6] Press et al. [1989] szerint tlagosan a Quicksort a leggyorsabb rendezsi algoritmus a legtbb gpen, nagy szm elemre, s $N \log N$ nagysgrendű a futsi ideje. Ugyanakkor kihangslyozzk a szerzők, hogy a legrosszabb esetben igazbl csak N^2-es az algoritmus.

gramot is (`planmk`), amely egy, a vektor-fjlokat tartalmaz listbl elkszt egy olyan vgrehajthat fjlt, amely a `szorzas` program sokszori meghvsval mindegyik vektort sszeszorozza az sszes 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 nevt, amellyel szoroznia kell a mozg doboz tartalmbl kpezett vektort. Ez a program a `vektor` s a `szorzs` program lnyeges elemeit egyarnt tartalmazza, hiszen minden lpsben a doboz tartalmbl vektort kell kpeznie, majd azt sszeszorozni a szorz-fjllal. A program olyan formtumban menti el az eredmnyeket, hogy az a `grapher` nevű adatfeldolgoz program szmra hasznlhat legyen.

A korrelcimentes ("nav") vektorokat a `naivvekt` program ksztette el (ld. a 4.5 fejezetet), amely bemenő paramterknt felhasznlta a *GenBank* fjljaiban szereplő bzisgyakorisgokat. 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 sszeszmolja a dobozban lvő szimblumokat fajtik szerint.

4.4 A felhasznlt adatbzis

Ahhoz, hogy Damashek-fle vektortrtechnikt DNS-szekvencikra alkalmazhassam, a `http://www.ncbi.`
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 feldolgozsra, de mg egy jelentős rsznek a feldolgozsra sem vllalkozhattam. Az albbiakban bemutatand eredmnyek teht egy meglehetősen szűk szekvencia-halmazra plnek, de arra utalnak, hogy rdemes a munkt folytatni, esetleg biolgiailag is motivlt krdsek 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 kromoszmjt (egy rgebbi vltozatban `SCCHRIII.SEQ` a III. kromoszmt jelli, ill. az jabban az sszes kromoszma `yeast_chr_`n`.gbk` nv alatt tallhat meg, ahol $n = $ `01, 02, ...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, bibliogrfiai adatokat, a ngy bzis arnyt, valamint az egyes rszletekkel kapcsolatos is-

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

4.3. tblzat: A NCBI-GenBank tartalma 1998. februr 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 elhelyezkedse, 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 tartalmaznak kdol szakaszokat (CDS-eket). Az lesztő kromoszmi kzl nhny nem tartalmaz intront, de tbbjkben tbb is tallhat, amelyek kzl a leghosszabbakat sszegyűjtttem, akrcsak nyolc hossz CDS-t (ld. a 4.4. tblzatban). A `yex2` exont tartalmaz fjlhoz 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, kztk kt intronnal; utbbiak kzl a msodik jval nagyobb valamennyi emltett szakasznl; vgezetl a

4.4. tblzat: Az lesztő genomjnak felhasznlt rszletei. A nyolc darab exont `yex`*n, mg a nyolc darab intront* `yin`*n jelli (*$n = 1..8$*).*

4.5. tblzat: A `HUMHBB.SEQ` *primary transcript-jeinek szerkezete: a vezrszakaszt kveti a hrom exon, kztk 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. tblzat: Az lesztő nhny exonjnak s legnagyobb intronjainak hasonlsga, $n = 4$ mellett. A nyolc exon exonbl ksztett vektorok szorzata $0,87 \pm 0,11$, az exonok s az intronok hasonlsga $0,76 \pm 0,095$, mg az intronok egyms kztt csupn $0,73 \pm 0,043$ hasonlsgot mutatnak.

szekvencik Damashek-fle vektorait, klnbző n-ek mellett, majd elvgeztem a megfelelő szorzsokat. Tekintsk t az rdekesebb eredmnyeket.

Az lesztő genomjbl kiemelt, a 4.4. tblzatban felsorolt exonok s intronok szorzatait mutatja a 4.6. tblzat, amikor $n = 4$-eseket hasznltam. A tblzat szerint meglehetősen magas az exonok egyms kztti hasonlsgi mrtke $(0,87 \pm 0,11)$. Ez az rtk mg magasabb lenne, ha eltekintennk a `yex2` exontl, amely, mint rtam, nagy mennyisgben tartalmaz ismtlődő szekvencikat. Az exonok s az intronok hasonlsga mintegy egy szrsnyival, kzel 15%-kal kevesebb, $0,76 \pm 0,095$. A kt adathalmaz egyrtelműen elklnlne, ha kihagynnk a `yex2` exont. Ezzel szemben, az intronok egyms kztti hasonlsga nem jelentős, csupn $0,73 \pm 0,043$. Figyelemre mlt viszont a kis szrs.

A HUMHBB s a RATCRYG-szekvencik mindegyik *primary transcript*-jből kt "szveget" ksztettem: az egyik a hrom exon sszefűzse, mg a msik a vezrszakasz, a kt intron s a vgszakasz, azaz a fehrjt nem kdol rszek konkatencija. A 4.7. tblzat a HUMHBB-szekvencinak a 4.5. tblzatban felsorolt t *primary transcript*-jből, $n = 3$ mellett generlt, kdol (E1-E5) s nem kdol (I1-I5) szakaszokat reprezentl vektorainak a szorzatt mutatja. A vizsglt tz szekvencia lesen kettvlik, hiszen az exonok egymssal $0,94 \pm 0,016$ hasonlsgot mutatnak, a nem kdol szakaszok egymssal val hasonlsga $0,92 \pm 0,03$, ugyanakkor a kdol s a nem kdol szakaszokbl ksztett vektorok szorzata csupn $0,75 \pm 0,08$. Figyeljk meg a kis szrst az első kt esetben, ami azrt is meglepő az exonok esetben, mert ezek rvid szekvencik. Hasonl eredmnyeket kaphatunk ms 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-jbl 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 komponenst gy kapom meg, hogy sszeszorzom az i-ik bzis-n-est alkot bzisoknak az adott szekvenciban megfigyelhető frekvenciit. Azaz, ha az i-ik n-gramm (bázis-n-es) alakja $\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}$, és ν_j jelöli a σ_j bázispár megfigyelt előfordulási frekvenciáját a vizsglt szekvenciban, akkor az $\mathbf{x}^{(naiv)}$ vektor i-ik komponense:

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

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 magyarzhat trivilis mdon azzal, hogy a *Ratus norvegicus* gnjből kpeztk a korrelcimentes vektort, hiszen az exonok nagyon kis hnyadt jelentik az egsz gn mretnek.

Ezekre az eredmnyekre 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 szorz fjl az egsz HUMHBB-szekvencia alapjn ksztett 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 vgig a szekvencin, s lpsről lpsre brzoltam a doboz tartalmbl kpezett Damashek-fle vektornak az egsz HUMHBB-szekvencibl szmtott nav (korrelcimentes) vektorral kpezett szorzatt, a doboz pozcijnak 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 gn delta-globulint kdol rszt hasonltottam ssze a HUMHBB egszből kpezett korrelcimentes vektorral. $n = 3$, $m = 200$. A pontozott pozcikban tartalmaz a doboz exon-rszletet: rdemes megfigyelni a kt első vlgy tlapolst.

a doboz nem angol nyelvű szvegbe lp. Így az exonokat egy vlgy jelzi a hasonlsg mrtkt a doboz kezdőpozcijnak a fggvnyben brzol diagramon: a doboznak minl nagyobb rsze tartalmaz exont, annl kisebb a szorzat rtke. Ha m a mozg doboz szlessge, akkor a vlgy $\frac{m}{2}$-vel balra van tolva az exonhoz kpest, hiszen a vlgy m bzissal az exon előtt elkezdődik, amikor a doboz egyik vge "belelp" az exonba, s az exon vgn fejeződik be, amikor a doboz msik vge is elhagyja az exont.

A három exon közül az első kettő nagyon közel van egymáshoz, így van olyan helyzet, hogy a doboz belelóg 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* ktharmada kzelben, mind az öt átírásra kerülő szakasz esetén ugyan ott. Ezen "ál-exon" létére még nem találtam magyarázatot.

Végül, a *4.4 ábrán* az élesztő III. kromoszómájának (SCCHRIII) egy részlete látható, az eddig bemutatott mozg dobozos technikval. Ugyan ezt a részletet mutatja be Stanley et al. [1993b] 3. ábrája (a *DNA-walk* α-exponense egy $m = 800bp$ széles mozgó dobozban). Mindkét ábrában lokális minimumok jelzik az exonok helyét (ld. a 3.3. fejezet vgn rtakat), 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 egsz kromoszma alapjn kpezett nav vektor. ($n = 3$)

5. fejezet

sszefoglals

Dolgozatom clja azon mdszerek bemutatsa volt, amelyek segtsgvel DNS-szekvencik statisztikai jellemzőit vizsglni lehet. Mivel a genetikai kd ugyan olyan szimblumszekvencia, mint a termszetes 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-analzist vagy az informcielmleti redundancia mrst a termszetes nyelvekre alkalmaztk elsőknt, mg a hossz tv korrelcik kimutatsa "vletlen bolyongs" segts-

gvel előbb a genetikban trtnt meg.

A 3. fejezetben ismertetett mdszerek alkalmazsa a kvetkező eredmnyekre vezetett:

1. A kdol s a nem kdol DNS-szakaszok tbb tekintetben is eltrő viselkedst mutatnak. Viszont sok hasonlsgot fedezhettnk fel utbbiak s a termszetes nyelveken rott szvegek kztt.
2. A nem kdol szakaszok megfigyelt statisztikai tulajdonsgai nem rhatk le Markov-folyamatok segtsgvel. Ezrt javasoltam a sztochasztikus krnyezetfggetlen grammatikk alkalmazst mindkt tpus szimblumsorozat esetben, s ismertettem ennek a modellnek az alapfogalmait. Kiszmoltam a modell paramtereiből az egyes terminlisok előfordulsi valsznűsgt, amelyre a Zipf-trvny kapcsn lesz szksgnk. (Ezt a szmtst tudtommal mg nem vgeztk el, mivel ezt a sztochasztikus mdszert jelenleg ms clokra alkalmazzk.)

Ezt kvetően, egy olyan eljrst alkalmaztam DNS-szekvencikra, amelyet eddig csupn rott szvegekre hasznltak, ott viszont az algoritmus eredmnyesen vlasztja szt a klnfle szvegeket nyelv s tartalom szerint. Az eredmnyeim azt mutattk, hogy a mdszer az exonok s az intronok klnvlasztsra is alkalmas:

1. Az evolci klnbző szintjein ll fajoknl egyarnt megfigyelhető volt az, hogy az exonok egymssal szignifiknsan magasabb hasonlsgot mutattak, mint az exonok s az intronok.
2. Kt emlős fajtl szrmaz gnen is megmutattam, hogy az intronokban gyengbbek a rvid tv korrelcik, mint az exonokban.
3. A "mozg dobozos" technikval megvizsglva egy emberi gnt, az exonok jellegzetes vlgyeket produkltak. De ez a mdszer jelenlegi llapotban mg nem alkalmas az exonok megkeressre egy adott szekvenciban, mivel ms szakaszok is eredmnyezhetnek hasonl vlgyeket.

A mdszer előnye az egyszerűsge. Egyszerű egyrszt komputcisan, vagyis nem ignyel se nem nagy szmtsi kapacitst, se nem előzetes "trning"-et, szemben ms, pldul neuronhlzatokat alkalmaz eljrsokkal. Msrszt maga az algoritmus is jl ttekinthető, s ezrt knnyen magyarzhat az, hogy milyen tnyezőkre megy vissza a sikere. A mdszer a szekvencik hasonlsgt a karakter-n-esek gyakorisgbl szmtja:

• Írott szvegek esetn az eltrő gyakorisgok oka lehet az eltrő tma (azaz ms tartalmas szavak gyakoriak a kt szvegben) vagy az eltrő nyelv. Utbbi esetben a szintaxis, a fonolgia s a helyesrs egyarnt felelős a karakter-n-esek eltrő gyakorisgairt.

• DNS-szekvencik esetn pedig az ok abban (is) keresendő, hogy Clay et al. [1996] szerint adott fajnl eltrő az intronok s az exonok GC-szintje, amely tny mr nmagban is elegendő lenne az intronok s az exonok kimutatott klnbzősgnek a magyarzatra.

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

Mantegna et al. [1995b] "nyelv" alatt procedurk s / vagy utastsok sorozatt rti, amelyek egy jl definilt nyelvtan szerint vannak elrendezve. Utal arra, hogy az ismert *genetikai kdon* tl, amely csupn a proteinek elsődleges szerkezetről tartalmaz informcit, lehetsges, hogy ltezik 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 meghatrozza a kommunikci egsz rendszernek mgttes szerkezete. Ugyanakkor a cikk vge hangslyozza, hogy semmifle bizonytkot nem talltak a "genetikai nyelv" lte mellett.

Mantegna nyelvfilozfiai mlysgű nyelv-fogalmval 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 nyelvszeti rtelemben teszik nyelvv a DNS-szekvencikat.

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

Irodalomjegyzk

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. jliusban, Budapesten tartott Econophyiscs Workshop sorn 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.): Strukturlis magyar nyelvtan, 2. ktet: Fonolgia, Akadmiai Kiad, Budapest.

B. Krenn, Ch. Samuelsson [1996]: The Linguist's Guide to Statistics,
forrs: `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 fejlőds termodinamikja (szemlyes pldny)

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, Tanknyvkiad, 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, Akadmiai 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 koszelmlet a trsadalomkutatsban, szerk.: Fokasz Nikosz, Replika Kr, Budapest

Rvsz Gy. [1979]: Bevezets a formlis nyelvek elmletbe, Akadmiai 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 szletse 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, Akadmiai 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.