

1.

Az információ fogalmköre,
adatbázisok/adatbankok szerepe az
orvosi gyakorlatban és a kutatásban

2014.12.01.

gp.

2.

A orvosi-/bio-informatika mint interdiszciplina

- ✓ BIOLÓGIA
- ✓ BIOTECHNOLÓGIA
- ✓ FEJLŐDÉSTAN
- ✓ FIZIOMIKA (EGY ÉLŐ SZERVEZET GENOMJÁTÓL A SZERVEZET EGÉSZÉIG TERJEDŐ KAPCSOLATOT ÍRJA LE)
- ✓ GENOMIKA
- ✓ INFORMÁCIÓTECHNOLÓGIA
- ✓ MATEMATIKA
- ✓ MOLEKULAMODELLEZÉS
- ✓ PROTEOMIKA
- ✓ STATISZTIKA

2014.12.01.

gp.

3.

Témakörök:

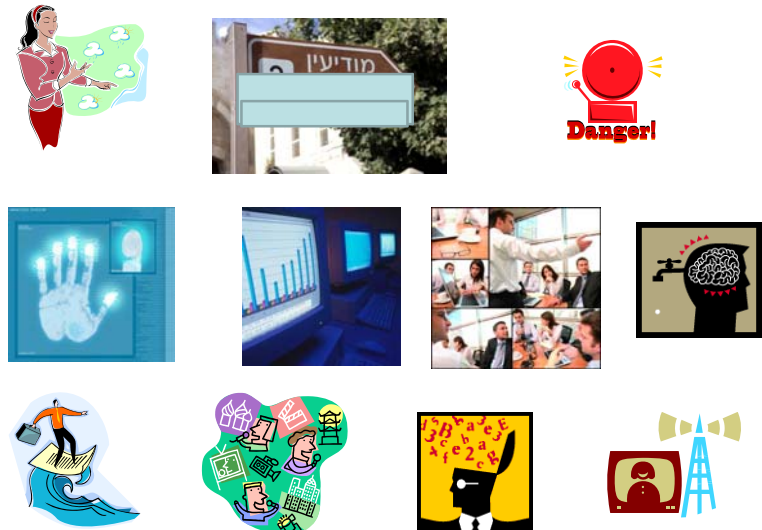
- I. Az információ fogalma, mértéke
- II. Kódolás. Kódolási hatások, redundancia
- III. A genetikai kód, információtartalma
- IV. Bioinformatikai adatbankok

2014.12.01.

gp.

4.

Az információ fogalma, mértéke



2014.12.01.

gp.

Mi az „információ”?



informatio:

későbbi latin értelmezés szerint: tanítás általi képzés; felvilágosítás, oktatás, tanítás a korábbi latin „előadás, magyarázat”-ból származtatva.

Információ:

- tudás, ismeret valamiről/valakiről;
- egy hír által közölt ismeret;
- adott helyzetről, folyamatról nyújtott/szerzett ismeret;

5.

2014.12.01.

gp.

Informatikai fogalomként:

Információ az a jelentés amit egy hír hordoz.



Az információ

- olyan új ismeret, ami a bizonytalanságot/határozatlanságot csökkenti.
- **jeleknek** olyan sorozata, elrendeződése amelyek meghatározott gyakorisággal lépnek fel;
- aminek jelentést tulajdoníthatunk;
- ami a címzettet egy meghatározott viselkedésre készíti

6.

2014.12.01.

gp.

Az információ – más definíció szerint - az "**értelemmel bíró adat**", és ennek megfelelően igen sokféle formában, különböző adathordozókon létezhet.

7.

Az információ-”szerzés” folyamata

hír tanulás
kommunikáció adatgyűjtés



jeleknek olyan sorozata, elrendeződése amelyek meghatározott gyakorisággal lépnek fel;

2014.12.01.

gp.

- jelek (a példákban)
- ✓ hangok, szavak, hanglejtés;
 - ✓ betűk, szavak, mondatok, kontextus
 - ✓ fiziológiai állapotot leíró jellemzők/jelek

8.

Az információ forrása, tárolása:



az "értelemmel bíró adat" igen sokféle formában, különböző adathordozókon létezhet

2014.12.01.

gp.

tárolás (pl.):

számítógépeknél:

- ✓ mágneses tárolók,
- ✓ optikai tárolók,
- ✓ integrált áramkörök (ROM, RAM, stb.)
- ✓ stb.

páciens esetén:

- az elsődleges forrás a beteg;
- a kapott adatok tárolása különböző módon valósul meg.

2014.12.01.

gp.

Számrendszerek:

tízis: 0,...,9; kettes: 0,1

$$2008_{(10)} = 2 \cdot 10^3 + 0 \cdot 10^2 + 0 \cdot 10^1 + 8 \cdot 10^0$$

$$2008_{(10)} = ?_{(2)} \longrightarrow$$

$2^0=1$	maradék	kitevő(n)	2^n	szorzótényező
$2^1=2$	2008	10	1024	1
$2^2=4$	984	9	512	1
$2^3=8$	472	8	256	1
$2^4=16$	216	7	128	1
$2^5=32$	88	6	64	1
$2^6=64$	24	5	32	0
$2^7=128$	24	4	16	1
$2^8=256$	8	3	8	1
$2^9=512$	0	2	4	0
$2^{10}=1024$	0	1	2	0
$2^{11}=2048$	0	0	1	0

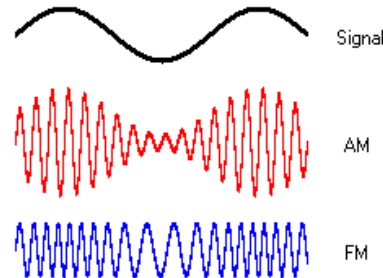
2014.12.01.

gp.

$$2008_{(10)} = 11111011000_{(2)}$$

**1 bit: egyetlen hely(érték) a számítógépes tárolásban;
1 byte: nyolc bit**

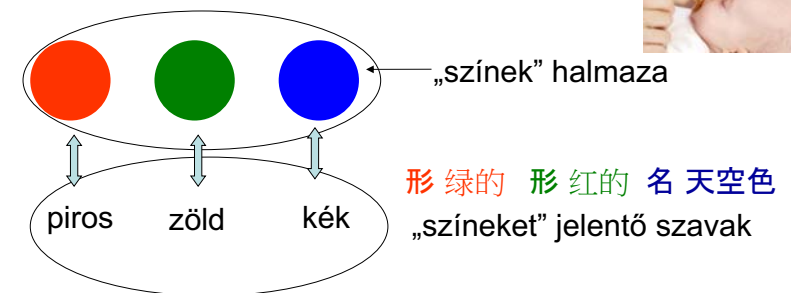
SI: 1kbit=10³ bit; (gyakran) számítástechnika: 1kbit = 1024 bit



2014.12.01.

gp.

Kódolás —dekódolás



kölcsönösen egyértelmű megfeleltetés két halmaz elemei között

adó: információt tárol/küld kódolt formában

vevő: információt fogad, dekódol

2014.12.01.

gp.

A kódolás feladata/szerepe

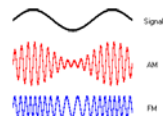
13.

- ✓ az információ tárolása, továbbítása egy adott jelrendszert alkalmazva

pl. Morse-kód
feromonok
DNS-szekvencia
hologramm

Jelrendszer:

adatok;
számok;
jelek (pl. piktogrammok, hieroglifák);
betűk;
aminosavak (fehérjék felépítésében);



feltétel:

- ✓ megegyezés az információ megfogalmazásában, a szabályokban az „adó” és a „vevő” között (pl. a „kék” ugyanazt jelentse; a múlt idő jele a „t”);
- ✓ a jel(hordozó) készletet mind az „adó”, mind a „vevő” ismerje;

2014.12.01.

gp.



Összefoglalás I.

14.

Információ — kódolás

- ✓ egy jelenségnek, tulajdonságnak adott jelrendszeren (kódolás) alapuló leírása, tárolása, továbbítása;
- ✓ feltételezve az „adó” és a „vevő” egyidejű vagy egymásutáni jelenlétét (információ átadás/áramlás) ↔ információ önmagában nem létezik

2014.12.01.

gp.

15.

Kérdések:

- Mekkora egy információ információtartalma?
- Hogyan lehet hatásosan kódolni?
- Hogyan lehetne általánosan leírni az információ továbbítását?



2014.12.01.

gp.

Információtartalom

16.

- a páciensnek egy foga lyukas
- a páciensnek minden foga lyukas

Megérzés:

a kisebb valószínűségű esemény információtartalma nagyobb

Mekkora egy továbbított üzenet információtartalma?

Információelméleti definíció:

Az információ a jeleknek olyan sorozata, elrendeződése amelyek meghatározott gyakorisággal lépnek fel.

↓
alma ↔ aalm

az információtartalom ugyanaz

2014.12.01.

gp.

Statisztikailag független események információtartalma

Legyen p : az adott jel (esemény) kimenetelének valószínűsége

A jelhez kapcsolódó információtartalom, $I(p)$

Definíció 1.:

$$I(p) = \log_2 \left(\frac{1}{p} \right) = -\log_2(p) \quad [I] = \text{bit vagy sh}$$

sh: Claude Shannon, az információelmélet megalapozója

Definíció 2.:

Azon biteknek a minimális száma (az információtartalom I , shannon egységeiben), ami ahhoz szükséges, hogy egyetlen, p -valószínűséggel fellépő jelet kódoltan, minimális jelkészlettel — **hatásosan** — továbbítsunk:

$$I(p) = -\log_2(p) \quad [I] = \text{bit v. sh}$$

Minél kisebb a jel előfordulási valószínűsége, annál nagyobb az információtartalma



$I(p=1) = -\log_2(1) = 0$ \longrightarrow ha az „üzenet” csak egyetlen jelből áll, akkor annak az információtartalma nulla.

pl.: legyen egy jel előfordulási valószínűsége $p=0,0625$.
Hány biten kell kódolni a jelet a maximálisan hatásos továbbításért?

$$I = -\log_2(0,0625) = 4 \text{ bit}$$

Kísérletsorozat (jelsorozat) információtartalma

m : osztályok száma (m -féle kimenetel; pl. ábécé betűi, kockadobás kimenetele 1—6, stb.)

p_k : a k -ik esemény valószínűsége/relatív gyakorisága

N : az összes esemény száma ($= n_1 + n_2 + \dots + n_m$; gyakoriságok)

Definíció 3.:

$$I = \sum_{k=1}^m n_k I_k = -\sum_{k=1}^m [n_k \cdot \log_2(p_k)]$$

Továbbítandó: „halandzsa” $I=?$; hány bit kell minimálisan a továbbításhoz?
Hogyan kódoljunk/tároljunk hatásosan?

Továbbítandó/tárolandó: „**halandzsa**”

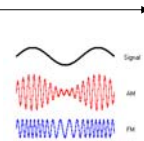
N=7 (!DZS!)

m=5



	$n_k(\text{gyak.})$	$f_k(\text{rel.gyak.})$	$I_k = -n_k \cdot \log_2(f_k)$
a	3	0.429	3.67
h	1	0.143	2.81
l	1	0.143	2.81
n	1	0.143	2.81
dzs	1	0.143	2.81
N=	7	$\Sigma I =$	14.90
		bitek száma	15

halandzsa



A hatásos kódoláshoz/tároláshoz 15 bit elegendő (a fenti példához!)

2014.12.01.

gp.

Biostatistika

2. A biostatistika szerepe, feladatai, leíró statisztika: az adat fogalma, adattípusok, az adatgyűjtés, az adatok ábrázolása, táblázatos ábrázolás, grafikonok.

3. A valószínűségi számítás elemei, a valószínűségi számítás és a statisztika kapcsolata (független események, feltételes valószínűség, esélyérték, esélyarány).

betű	gyak.	betű	gyak.	betű	gyak.
a	36	j	1	s	18
á	13	k	11	sz	10
b	5	l	14	t	25
c	0	ly	2	ty	0
cs	1	m	5	u	1
d	5	n	5	ú	0
e	18	ny	2	ű	1
é	11	o	11	ű	4
f	5	ó	4	v	3
g	7	ö	0	x	0
gy	1	ő	0	y	0
h	0	p	3	z	6
i	10	q	0	zs	0
í	7	r	7		

n=252

!dz,dzs!



2014.12.01.

gp.

Hogyan lehet hatásosan kódolni?

Cél:

- ✓ tárolás
- ✓ továbbítás
- ✓ legkisebb befektetéssel (energia, idő)



Megoldás:

1. Az információ információtartalmának megfelelően (minimálisan szükséges bitek száma)
2. A nagyobb gyakorisággal előforduló jelekhez a „legegyszerűbb/legrövidebb” kód hozzárendelésével.

2014.12.01.

gp.

A redundancia szerepe

Redundanciának nevezzük azt a jelenséget, amikor egy jelsorozatban egyes jelek előfordulását korábbi, vagy későbbi jelek alapján meg lehet jósolni.

pl.:

- „q” utáni „u”
- személyi/TAJ számokban az utolsó jegy

Következmény:

- ✓ kisebb az információátvitel hatásossága
- ✓ lehetőség a dekódolás/az átvitel minőségének ellenőrzésére/javítására (zajos az átviteli csatorna; pl.: igen gyöngye fényben nézünk valamit,...)

2014.12.01.

gp.

Az átlagos információtartalom



$$\bar{x} = \frac{\sum x_i}{N}$$

Definíció 4.:

$$\bar{I} = \frac{\sum_{k=1}^m n_k \cdot I_k}{N} = -\sum_{k=1}^m \left[\frac{n_k}{N} \cdot \log_2(p_k) \right] = H$$

H: a kísérlet/jelsorozat entrópiája; egysége bit

$$H = \bar{I} = -\sum_{k=1}^m [p_k \cdot \log_2(p_k)]$$

2014.12.01.

gp.

A genetikai kód, információtartalma



Kérdések:

- mennyi a minimális jelkészlet ~hús aminosav kódolásához?
- mennyi egy DNS-szekvencia információtartalma?

Válasz 1 (lásd biológiai előismeretek is):

- ✓ négy nukleotid kódol
- ✓ ha párban kódolnának akkor a lehetséges variáció: $4^2=16$
- ✓ ha tripletben: $4^3=64$
- a tripletben való kódolás elegendő és minimális
- vannak aminosavak amit több kodon is kódol
- vannak különleges triplettek, amik más funkciót irányítanak

2014.12.01.

gp.

Válasz 2:

Tf.: azonos valószínűséggel fordulnak elő a bázisok

— $p_k=p=0,25$; $I_1=I_2=I_3=I_4=I_b$

Ha a szekvencia hossza N, akkor $n_k=n=N/4$

$$I = \sum_{k=1}^4 n_k I_k = nI_1 + nI_2 + nI_3 + nI_4 = 4 \cdot n \cdot I_b$$

$$I = 4 \cdot N / 4 \cdot I_b = N \cdot I_b = -N \cdot \log_2(p)$$

$$I = -N \cdot \log_2(0,25) = N \cdot 1,6021 \text{ bit}$$

$$N=10 \rightarrow \sim 16 \text{ bit}$$

$$N=10^6 \rightarrow \sim 1,6 \cdot 10^6 \text{ bit}$$

2014.12.01.

gp.

Összefoglalás II.



- Egy információ információtartalmát a Shannon által bevezetett (információs) entrópiával jellemezhetjük;
- Egy információs csatornára vonatkozó maximális kódolási hatásfokot az információtartalomnak megfelelő minimális kóddal érhetjük el.
- Fehérjék, DNS vagy más makromolekula által hordozott információt az aminosavak/bázisok/monomerek gyakorisága alapján számíthatjuk.

2014.12.01.

gp.

Bioinformatikai adatbankok



Cél:

a biológiai, orvosi gyakorlatban megszerzett ismeretek

- tárolása,
- rendszerezése,
- minőségellenőrzése,
- analizálása,
- elérhetővé tétele

Követelmény:

- ✓ gyors, hatékony hozzáférhetőség;
- ✓ csak azoknak az információknak a kinyerése, amik az adott felhasználót érdeklik.

2014.12.01.

gp.

Követelmény:

gyors, hatékony hozzáférhetőség =

= csak azoknak az információknak a kinyerése, amik az adott felhasználót érdeklik.

Specializált adatbankok:

előny: rövidebb találati idő, részletes adatok

hátrány: az összefüggések hiánya

Kevésbé specializált adatbankok:

előny: adatok/jelenségek közötti összefüggések kereshetők

hátrány: több szempont szükséges adott ismeret megtalálásához

2014.12.01.

gp.

Kereső algoritmusok (kereső motorok):

az adott adatbázisban keresett ismeret megtalálását célzó matematikai/informatikai eljárás

Gyakorlati tudnivalók az adatbázisokban való kereséshez:

- 1.) „Józan paraszti ész”: — a keresett fogalomnak megfelelő adatokat tartalmazó adatbázisban keressünk
- 2.) túl általánosan definiált kérdés — túl sok eredmény
- 3.) túl speciálisan feltett kérdés — túl szűk eredményhalmaz
- 4.) a második módszert első kereséskor használjuk, a harmadikat „majdnem ismert” válasz esetén.

pl.: cholesterol — 182358
 cholesterol transport — 9055
 cholesterol transport pediatrics — 128

2014.12.01.

gp.

cholesterol transport pediatrics Chan T. — 2

Jelinek D, Patrick SM, Kitt KN, **Chan T**, Francis GA, Garver WS.: Physiological and coordinate downregulation of the NPC1 and NPC2 genes are associated with the sequestration of LDL-derived cholesterol within endocytic compartments. J Cell Biochem. 2009 Sep 10. [Epub ahead of print] PMID: 19746448

Sahoo D, Trischuk TC, **Chan T**, Drover VA, Ho S, Chimini G, Agellon LB, Agnihotri R, Francis GA, Lehner R. ABCA1-dependent lipid efflux to apolipoprotein A-I mediates HDL particle formation and decreases VLDL secretion from murine hepatocytes. J Lipid Res. 2004 Jun;45(6):1122-31. Epub 2004 Mar 1.

2014.12.01.

gp.

33.

GenBank from NCBI (National Center for Biotechnology Information) Genetic Sequence Databank;

EMBL Nucleotide Sequence Database (European Molecular Biology Laboratory);

SwissProt és **PROSITE** (protein sequence database);

EC-ENZYME (a már jellemzett enzimek adatbankja);

RCSB PDB (3-D biológiai makromolekuláris szerkezetek Rtg-diffrakció-, NMR-, and Cryo-EM alapján);

MEDLINE: humán medicina, fogászat, állatorvosi tudomány, kísérletes orvostudomány,...

PUBMed (<http://www.ncbi.nlm.nih.gov/sites/entrez>): bibliográfiai adatbázis orvostudomány, biológia, biokémia, biofizika,...

EISZ (<http://www.eisz.hu>): magyarországi főiskolák és egyetemek oktatói, hallgatói számára hozzáférés (internetcím alapján).

2014.12.01.

gp.

34.

SOTE-n belül:

<http://www.lib.sote.hu/>

Források

Adatbázisok

Tudományos cikkek

Könyvek

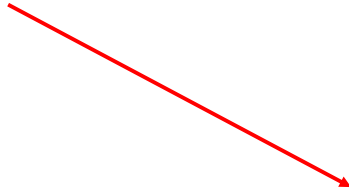
Tudományometriai adatbázisok

Gyógyszerészeti adatbázisok

2014.12.01.

gp.

35.



C. Shannon (1916-2001)

2014.12.01.

gp.