Grundbegriffe der Informatik
Rolle/Funktion der medizinischen Datenbanken in Praxis und Forschung

Medizinische-/Bio-Informatik als inter-(multi-)disziplinäres Gebiet

- ✓ ***Biologie***
- ✓ ***Biotechnologie***
- ✓ ***Entwicklungslehre***
- ✓ ***Physiomik* (ab Niveau der Gene bis Zusammenfunktion von Geweben, Organen)***
- ✓ ***Genomik****
- ✓ ***Informationstechnologie***
- ✓ ***Mathematik***
- ✓ ***Molekülmodellierung***
- ✓ ***Proteomik***
- ✓ ***Statistik***

*Empfohlen als wertvolles zu lesen
http://www.uni-heidelberg.de/presse/news/2106bartram.html

Themen:

I. Begriff und Maß der Information
II. Codierung. Wirkungsgrad, Redundanz
III. Genetischer Code, sein Informationsgehalt
IV. Bio-Datenbanken

Begriff und Maß der Information









Ein Zitat von Augustinus (*354 — †430):

„Was ist also *die Zeit*? Wenn mich niemand danach fragt, weiß ich es, wenn ich es aber einem, der mich fragt, erklären sollte, weiß ich es nicht."

„Was ist also *die Information*? Wenn mich niemand danach fragt, weiß ich es, wenn ich es aber einem, der mich fragt, erklären sollte, weiß ich es nicht."

Was ist die Information?

"informatio":
lat.: „bilden", „eine Form, Gestalt, Auskunft geben.

Information:

- Wissen (s. Wikipedia: von althochdeutsch *wizzan*; zur indogermanischen Perfektform *woida* "ich habe gesehen"), Kenntnis von jemandem/etwas;
- Kenntnis auf Grund Nachrichten;
- Kenntnis/Wissen von einem gegebenen Umstand/Prozess;

Information als Begriff der Informatik:
Information ist diejenige Bedeutung, welche durch eine Nachricht getragen ist.

weitere Definitionen:

die Information:

- eine neue Kenntnis, die die Ungewissheit/Unbestimmtheit vermindert.
- ***Reihenfolge/Struktur der Zeichen,*** worin die Zeichen mit bestimmten Wahrscheinlichkeiten auftreten;
- ihr Bedeutung beimessen werden kann – besitzt Bedeutungsgehalt;
- treibt den Empfänger/Adressat zu einem bestimmten Verhalten, einer Bewegung an

Information enthält – auf Grund einer anderen Definition - "***sinnvolle Daten***";

- besitzt mehreren Formen;
- kann auf unterschiedlichen Datenträgern gespeichert werden, vorhanden sein.

Wie kann man Informationen erwerben?

Nachricht | Kommunikation | Studien | Datensammlung




Reihenfolge/Struktur der Zeichen, worin die Zeichen mit bestimmten Wahrscheinlichkeiten auftreten

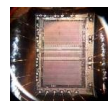
Zeichen (entsprechend der Bilder);

- ✓Stimme, Worte(Wörter), Klang/Intonation;
- ✓Buchstaben, Worte, Sätze, Kontext;
- ✓ die den physiologischen Zustand beschreibenden Eigenschaften, Charakteristiken

Quelle und Speicherung der Informationen

Als "sinnvolle Daten", kann die Information in unterschiedlichen Formen, auf verschiedenen Datenträgern gespeichert werden.

Speicherung (Z.B.):
bei Computern:
- ✓magnetisch,
- ✓optisch,
- ✓integrierte Schaltkreise (ROM, RAM,...)
- ✓usw.:



in medizinischer Praxis:
- die primäre Quelle ist der Patient;
- Speicherung der gewonnenen diagnostischen Testwerte;

Zahlensysteme:

Dezimal: 0,...,9; Binär: 0,1

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

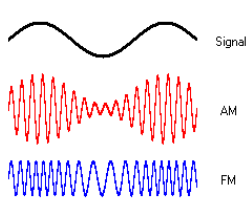
	Rest	Potenz(n)	2^n	Multiplikator
2^0=1	2008	10	1024	1
2^1=2	984	9	512	1
2^2=4	472	8	256	1
2^3=8	216	7	128	1
2^4=16	88	6	64	1
2^5=32	24	5	32	0
2^6=64	24	4	16	1
2^7=128	8	3	8	1
2^8=256	0	2	4	0
2^9=512	0	1	2	0
2^{10}=1024	0	0	1	0
2^{11}=2048				

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

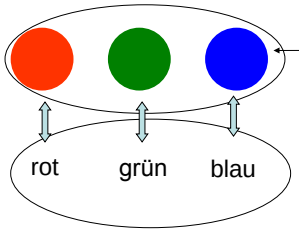
1 bit: eine einzige Stelle für Datenspeicherung in Computern (Taschenrechnern);
1 byte: acht bit

SI: 1kbit=10^3 bit; (meistens) in Informatik: 1kbit = 1024 bit

Code



Codierung – Decodierung (Ver/Ent-schlüsselung)



形 绿的 形 红的 名 天空色

Wörter für Bezeichnen eine Farbe

gegenseitig-eindeutige Zuordnung zwischen den Elementen zwei Mengen

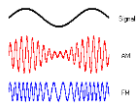
Sender: sendet/speichert /codiert Informationen in verschlüsselter Form;
Empfänger: empfängt und entschlüsselt die übertragenen Informationen

Die Rolle und Funktion der Codierung

✓ Speicherung und Übertragung der Informationen durch Anwendung ein bestimmtes Zeichensystem

z.B.: Morse-Code
Pheromone
DNS-Sequenzen
Hologramm

Zeichensysteme:
Daten;
Zahlen;
Zeichen (z.B: Piktogramme, Hieroglyphen);
Buchstaben;
Aminosäuren (im Aufbau von Proteinen);



Bedingungen:
- ✓ Vereinbarung zwischen dem Sender und Empfänger in den Formulierungen und Regeln (eingeschlossen die Übertragungsmethode) der Informationen z.B.: die Zeichenfolge "blau" muss zweiseitig das selbe bedeuten;
- ✓ die Zeichensätze müssen für den Sender und den Empfänger bekannt sein;

Zusammenfassung I.

Information — Codierung
- ✓ *Beschreibung, Speicherung und Übertragung einer Erscheinung, Eigenschaft mit Hilfe eines Zeichensystems (Codierung);*
- ✓ *angenommen, dass der "Sender" und der "Empfänger" gleichzeitig oder nacheinander anwesend sind (Informations-übertragung/fluss) ↔ die Information existiert nicht allein/selbstständig*

Fragen

- ➢ Wie groß ist der Informationsgehalt einer Information?
- ➢ Wie kann man effizient Codieren?
- ➢ Wie wäre es möglich den Informationsübertrag im Allgemeinen zu beschreiben?

Informationsgehalt

- der Patient hat einen lockeren Zahn
- alle Zähne eines Patienten sind locker

Welche dieser Informationen weist größeren Gehalt auf?

Auf Grund Gefühle, Eingebung:
eine Information mit geringerer Wahrscheinlichkeit weist größeren Informationsgehalt auf.

Auf Grund einer Definition der Information:

- *Reihenfolge/Struktur der Zeichen, worin die Zeichen mit bestimmten Wahrscheinlichkeiten auftreten*

Affe ⟺ Effa

derselbe Informationsgehalt

Informationsgehalt der statistisch unabhängigen Ereignissen

Bezeichne ***p*** die Wahrscheinlichkeit eines gegebenen Ereignisses (d.h. jetzt Zeichen)

Der Informationsgehalt, I(p), dieses Zeichens ist:

Definition 1.:

$$I(p) = \log_2\left(\frac{1}{p}\right) = -\log_2(p)$$

[I]=bit oder sh

sh; nach dem Namen Claude Shannon, der Begründer der Informationstheorie

Definition 2.:

Der Informationsgehalt ist gleich der minimalen Anzahl der Bits (in sh-Einheiten), die benötigt sind um ein Zeichen der Wahrscheinlichkeit p, verschlüsselt mit minimalen Zeichensatz, effizient zu übertragen:

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

Beispiel:
- p = 0,5 I(p) = $-\log_2(0{,}5) = -\log_2(1/2) = \log_2(2) = 1$;
- p = 0,25 I(p) = $-\log_2(0{,}25) = -\log_2(1/2^2) = 2*\log_2(2) = 2$

Je kleiner die Auftrittswahrscheinlichkeit eines Zeichens ist, desto größer ist sein Informationsgehalt.

$I(p=1) = -\log_2(1) = 0$ ⟹ besteht die "Botschaft" nur aus einem einzigen Zeichen, ist sein Informationsgehalt gleich Null

Quiz: Wenn die relative Häufigkeit eines Zeichens p=0,0625 ist, wie viele Bits sind nötig für die maximal-effiziente Übertragung?

I=$-\log_2(0{,}0625) = -\lg(0{,}0625)/\lg(2)$=4 bit

Informationsgehalt einer Zeichenfolge

m: Anzahl der unterschiedlichen Ereignisse (m-Ausgänge eines Versuches; d. Anzahl der Buchstaben, usw...);
p_k: relative Häufigkeit/Wahrscheinlichkeit eines Ereignisses;
N: Anzahl der allen Ereignissen (= $n_1+n_2+...+n_m$; Häufigkeiten)

Definition 3.:

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

zu übertagen:
„abrakadabra"

I=?; wie viele Bits ist, minimal, benötigt zu einer Übertragung?
Wie kann man am effizientesten verschlüsseln, speichern?

N=1
m=5

Zeichen	n_k (Häufigkeit)	f_k (rel. Häufigkeit)	$I_k = -n_k * \log_2(f_k)$
a	5	0.4545	5.688
b	2	0.1818	4.919
r	2	0.1818	4.919
k	1	0.0909	3.459
d	1	0.0909	3.459
N=	11		
		ΣI= 22.444	

Eine effiziente Übertragung benötigt 23 bit für "Abrakadabra"

Wie kann man effizient kodieren?

Ziel:
der minimale Aufwand an Zeit, Energie,...
- ✓ Speicherung
- ✓ Übertragung

Lösung:
1. gemäß dem Informationsgehalt (d.h. mit minimaler Anzahl der benötigten Bits)
2. Zuteilung der kürzesten Codes zu Zeichen mit höchster Wahrscheinlichkeit.

die Rolle der Redundanz

Redundanz (nach Duden):
- das Vorhandensein von eigentlich überflüssigen, für die Information nicht notwendigen Elementen;
- Überladung mit Merkmalen

Redundanz ist diejenige Eigenschaft/Phänomen, wenn das Auftreten einiger Zeichen in einer Zeichenfolge, auf Grund früherer oder späterer Zeichen, vorhersagbar ist.

z.B.:
➢„q" ist verbunden mit einem folgenden Auftreten von „u"
➢in früheren Zeiten (Ehepaar = Frau und Man)

Konsequenzen:
- ✓ die Effizienz der Informationsübertragung nimmt ab
- ✓ Möglichkeit für Kontrolle/Korrektur/Reparatur der Kodierung/Übertragung (geräuschvolle Übertragungskanale, teilweise verlorene Signale, ...)

der durchschnittliche Infromationsgehalt – Entropie einer Information

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

Definition 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: Entropie einer Versuchsserie/Zeichenfolge; Einheit: bit

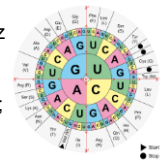
$$H = \bar{I} = -\sum_{k=1}^{m}\left[p_k \cdot \log_2(p_k)\right]$$

Informationsgehalt der genetischen Codes

Fragen:

- 1. wie groß sei der minimale Zeichensatz (Anzahl der unterschiedlichen Nukleotiden Codon) für Kodierung der ~ 20 Aminosäuren?
- 2. Wie groß ist der Informationsgehalt einer DNS-Sequenz?

Antworten 1.: (wie es bekannt is 4)



- ✓ Nukleotid*paare* ermöglichen nur einen Variationssatz 4^2=16;
- ✓ Nukleotidtripletts (Codons) – 4^3 = 64;
- ➢ Kodierung über Triplets ist minimal und ausreichend;
- ➢ einige Aminosäuren sind durch mehreren Codons kodiert;
- ➢ einige Tripletts besitzen andere Funktionen

Antwort 2:

- ✓ sei angenommen, dass die Nukleotidbasen treten gleichwahrscheinlich auf
- ✓ — p_k=p=0,25; I_1=I_2=I_3=I_4=I_b
- ✓ Wenn die Länge der Sequenz N ist, dann ist n_k=N/4=n

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

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

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

N=10 ⟶ ~16 bit

N=10^6 ⟶ ~1,6·10^6 bit

Zusammenfassung II.

- der Informationsgehalt einer Information kann mit der Informationsentropie beschrieben werden;
- der maximale Wirkungsgrad, die größte Effizienz, einer Kodierung wird mit einer dem Informationsgehalt entsprechenden minimalen Zeichensatz erreicht.
- Die durch DNS- oder Proteinmoleküle getragenen Informationen kann auf Grund der Häufigkeiten der monomer Nukleotiden oder Aminosäuren berechnet werden.

Bioinformatische Datenbanken

30.

Zielsetzung, Aufgabe:

- Speicherung,
- Organisierung,
- Qualitätskontrolle,
- Analyse,
- der Öffentlichkeit zugänglich machen

der Daten, Wissen, Kenntnisse der biologischen, medizinischen Wissenschaften

Anforderungen:

- ✓ schnell und effizient Zugriff;
- ✓ Auffinden nur für die Benutzer wichtige, wesentliche Informationen.

31.

spezialisierte Datenbanken:
Vorteil: kürzer Zugriffszeit, mehr detailierte Daten
Nachteil: Mangel an Zusammenhängen

weniger spezialisierten Datenbanken:
Vorteil: auch die Zusammenhänge zwischen den Daten/Erscheinungen sind durchsuchbar.
Nachteil: mehrere Gesichtspunkte sind gebraucht für Auffinden einer Kenntnis

Z.B.: cholesterol — 182358
cholesterol transport — 9055
cholesterol transport pediatrics — 128

32.

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.

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 ;
RCSB PDB (3-D makromolekularer Aufbau);
MEDLINE: Medizin, Zahnmedizin, Veterinärmedizin, forschungsmedizinische Informationen,...
PUBMed (http://www.ncbi.nlm.nih.gov/sites/entrez): Medizin, Biologie, Biochemie,...

Innerhalb der Universität:
http://www.lib.sote.hu/
Quellen
Datenbanken

Wissenschaftliche Artikel
Bücher
szientometrische/bibliometrische Datenbanken
pharmazeutische Datenbanken

C. Shannon (1916-2001)