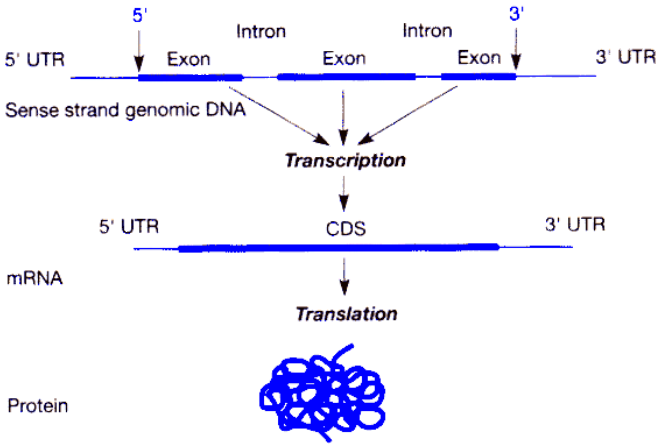
6. Nukleotidszekvenciák analízise

Bevezetés

Centrális dogma



Rövidítések: UTR: Untranslated Region, CDS: Coding sequence

A genetikai kód

	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
					C
	Leu		**Stop!**	**Stop!/Sec**	A
				Trp	G
C	Leu	Pro	His	Arg	T
					C
			Gln		A
					G
A	Ile	Thr	Asn	Ser	T
					C
			Lys	Arg	A
	Met/Start				G
G	Val	Ala	Asp	Gly	T
					C
			Glu		A
					G

A genetikai kód közel univerzális (apróbb eltérések azért vannak egyes mikroorganizmusokban, mitokondriumokban, stb.)

A TGA idõnként nem Stop−ot, hanem szelenociszteint (Sec) kódol (21. aminosav!)

DNS−szekvenciák összeszerelése

- cDNS szekvenálásánál 200−500 bázis hosszú darabokat kapunk (egy menetben ekkorát lehet szekvenálni, újabban azért max. 1000−et is)
- Szekvenálási hibák: kb. 5% (hibás bázisok; kimaradt/tévesen beszúrt bázisok: ún. fantom INDEL−ek)
- Emiatt mindkét szálat többször meg kell szekvenálni a megbízhatóság végett
- A teljes szekvenciát a darabokból kell összeszerelni és konszenzust konstruálni:

```
                AACCGTTTACGAAACCAGGTGC
                AACCGTTTACGAACCAGGTGCGCGCCCGCGGGAAT
                AACCGTTTACGAACCCAGGTGC
(konszenzus:)   AACCGTTTACGAAaCCAGGTGCGCGCGCGcGGGAATCCTAAAAA
                                    CGCGCGCGCGGGAATCCTAAAAA
                              AGGGAATGCGTAAAAAG
```

Kisbetûk: kisebb megbízhatóság

- Összeszerelés: különféle programokkal, pl. TIGR Assembler

Génvadászati forgatókönyv

1. Kiszûrni a repetitív elemeket (kiderül, hol *nem* lehetnek a szabályozó és kódoló régiók
2. Hasonlósági keresést végezni az ismert szekvenciák adatbázisain
3. Megvizsgálni a kodongyakoriságok eltolódásait (a fehérjekódoló régiók legegyértelmûbb jele)
4. Funkcionális helyeket (iniciáció, termináció, exon−intron határok, promoterek, stb.) keresni az ismert szekvenciamintázatok alapján
5. A fenti lépésekbõl nyert információt egységes képbe foglalni, konszenzusos képet kialakítani

Fontos:

- A legtöbb program specifikus valamely fajra vagy élõlénycsoportra
- A programokat a megfelelõ kontextusban kell használni (pl. exon−intron határt ne keressünk cDNS−ben)

A repetitív elemek kiszûrése

- Repetitív elemek: rövid (vagy hosszabb) ismétlõdõ szekvenciadarabok
- Eukariota genomok jelentõs részét teszik ki
- "Interspersed" repeat: a genomban elszórva találhatóak, többnyire transzpozábilis elemek (pl. transzpozonok) inaktiválódott kópiái
- Számos családjuk van, adatbázisuk: RepBase
- Minden DNS−elemzés elõtt ki kell szûrni ezeket a szekvenciából, mert minden elemzõprogramot megzavarnának.
- Jelentõsége: megmutatja, hol *nem* lehetnek szabályozó és kódoló régiók
- Programok erre: pl. Censor, RepeatMasker: a RepBase vagy hasonló adatbázis alapján szûrnek
- Pl. Censor futtatás eredménye:

Bemenõ szekvencia: humán kreatin kináz génjének részlete

```
> HUMCKMM1
ggatccttcctccttggcctcccaaagtgctgggattacaggtgtgagccactgcacctg
gcctattacccttctcaggctctggagtccatccttctgctctgtctccctcagttcaat
tgtttttttgtttttttgtttttttttttagacacagtctcgctctgtcaccaaggctggagt
gcagcagtgcgatcacagctcaccgcagcctcacctcccaggctcaagtgatcctcccat
ctcggcctctgagtagctgagactataggtgtgtccacatgtccggctaatttttgtatt
tttagtagagacagggtttcaccgcgttggccagggtggtcttgaactcctgagctcaag
caatcctcctgcctcagcctccttgttttgatttttagatcccacaaataacttgtgatg
tttgtctttctatacctggttcatttaacatttttctttttctttttctttttttttt
ttttttgtgagactgagtcttgctctgtcactcaggctggagggcaatggtgcatctcag
ctcactgcaacctccacctcctaggttcaagcaattcttatgcctcagcctcctggctag
ctgggattacaggcgtgtgtcaccatgccaggctaatttttgtacttttagtagagatgg
ggtttcaccatgttggccaggctggtcttgaactcctggcctcaagtgatccacccgcct
ccgcctctgcctcccaaagtgctgggattacgggcctgagccactgtgcccggcccatct
aacattttcactgtcaatcacaatgggattaaaactcctcccacagcccctagggacca
```

```
CENSOR futtatás eredménye:

Megtalált repetitív elemek:

            kezdet    vég   elem neve
humckmm1         2     63   Alu-Jb           1    62 c
humckmm1        67    119   L1MA2          697   751 c
humckmm1       138    382   Alu-Jb          42   290 c
humckmm1       383    449   L1MA2          623   696 c
humckmm1       451    480   (TTTTC)          5    33 d
humckmm1       481    775   Alu-Sz           1   290 c

A repetitív elemektõl megtisztított szekvencia (a törölt részek kiikszelve):

GXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXTATXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXTTGTTTTTTGTTTTTTGTXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXCATCTAACATTTTCACTGTCAATCACAATGGGATTAAAACTCCTCCCACAGCCCCTAGGGACCA
```

Hasonlósági keresések

A génazonosítás legjobb módszere, ha találunk a szekvenciaadatbázisokban már meglévõ, hasonló szekvenciát. Ennek módjai:

- A szekvencia összehasonlítása a GenBank−kal, ill. a dbEST−vel (BLASTN)
- A szekvencia mind a 6 lehetséges konceptuális transzlációjának összehasonlítása a fehérjeszekvencia−adatbázisokkal (BLASTX)
- A szekvencia transzlációinak összehasonlítása a GenBank transzlált változatával (TBLASTX)
- A szekvencia transzlációinak összehasonlítása másodlagos fehérjeszekvencia−adatbázissal (pl. PROSITE)

Csapdák:

- Esetenként benne maradhat az mRNS−ben, így a cDNS−ben egy intron. Ekkor ezt tévesen exonnak vélhetjük.
- A fehérjeszekvencia−adatbázisban nem valódi fehérjeszekvenciák is lehetnek (szintén transzlációk)
- Az új szekvenciáknak csak kb. a fele mutat hasonlóságot valamilyen régebbi szekvenciával
- Stb.

ORF−detektálás

- **ORF**: Open Reading Frame ("nyitott leolvasási keret"): olyan nukleotidszekvencia−szakasz, melyet Stop kodon zár le.
- Melyik a valódi leolvasási keret? Ehhez elõállítjuk mind a 6 lehetséges transzlációt (a DNS mindkét szála leolvasható, egymással ellentétes irányban, és mindkét szál háromféleképpen olvasható le), pl.

```
Your Nucleic Acid Sequence:

              10         20         30         40         50
    0 TCCATTGAGC CTTATACCAG TAACATCTAC ACTCGAAGAT CTTGTCAGGG
   50 GAATTTCAGA TTGTGAATCC TCACTTACTG AAAGATCTTA CTGAGCGGGG
  100 CTTGTGGAAT GAAGAGATGA AAAATCAGAT TATTGCATGC AATGGCTCCA
  150 TTCAGTTTTC CTTTTTCAGA GCATACCAGA AATTCCTGAT GACCTGAAGC
  200 AACTCTATAA GACCGTGTGG GAAATCTCTC AGAAGACTGT TCTCAAGATG

Six−Frame Amino Acid Translation:

 Forward 0:

             10         20         30         40         50
   0 SIEPYTSNIY TRRSCQGNFR L!ILTY!KIL LSGACGMKR! KIRLLHAMAP
  50 FSFPFSEHTR NS!!PEATL! DRVGNLSEDC SQD
```

```
Forward 1:

            10         20         30         40         50
  0 PLSLIPVTST LEDLVRGISD CESSLTERSY !AGLVE!RDE KSDYCMQWLH
 50 SVFLFQSIPE IPDDLKQLYK TVWEISQKTV LKM

Forward 2:

            10         20         30         40         50
  0 H!ALYQ!HLH SKILSGEFQI VNPHLLKDLT ERGLWNEEMK NQIIACNGSI
 50 QFSFFRAYQK FLMT!SNSIR PCGKSLRRLF SR

Reverse 0:
            10         20         30         40         50
  0 HLENSLLRDF PHGLIELLQV IRNFWYALKK EN!MEPLHAI I!FFISSFHK
 50 PRSVRSFSK! GFTI!NSPDK IFECRCYWYK AQW

Reverse 1:

            10         20         30         40         50
  0 ILRTVF!EIS HTVL!SCFRS SGISGML!KR KTEWSHCMQ! SDFSSLHSTS
 50 PAQ!DLSVSE DSQSEIPLTR SSSVDVTGIR LNG

Reverse 2:
            10         20         30         40         50
  0 S!EQSSERFP TRSYRVASGH QEFLVCSEKG KLNGAIACNN LIFHLFIPQA
 50 PLSKIFQ!VR IHNLKFP!QD LRV!MLLV!G SM
```

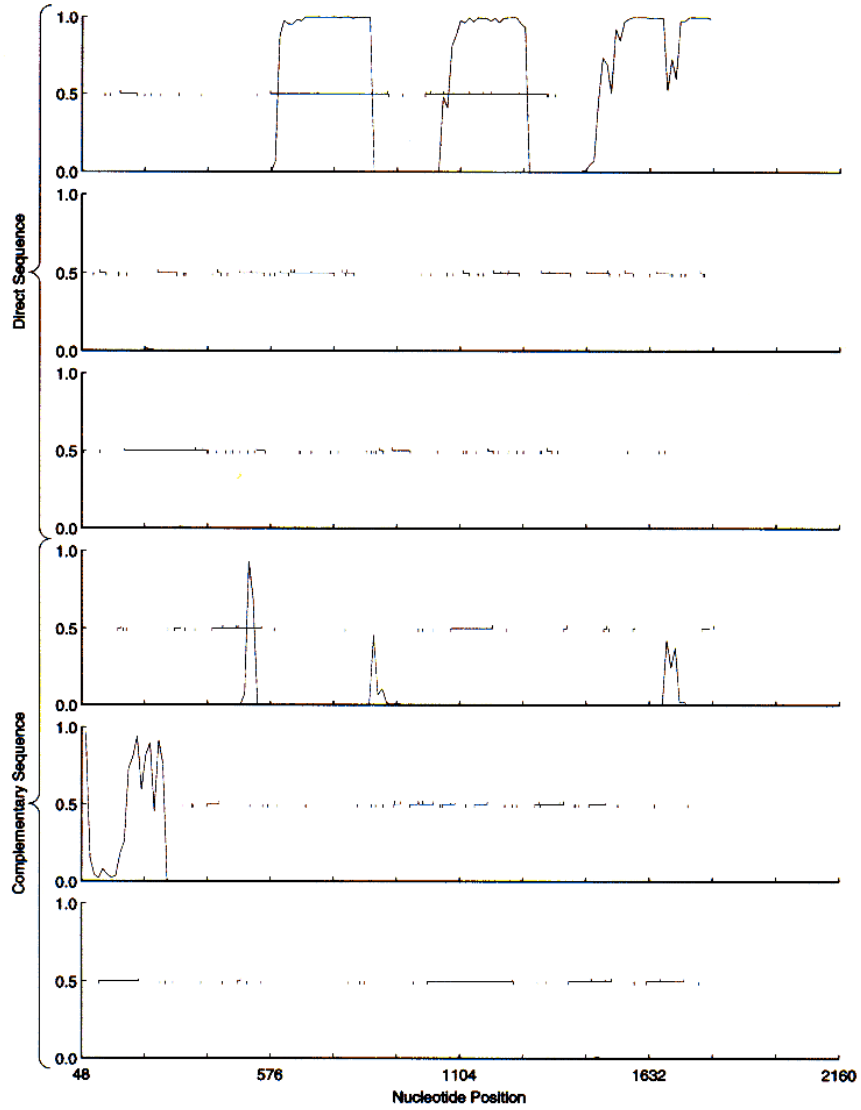
- A valódi leolvasási keret valószínûleg a leghosszabb megszakítatlan szekvenciát eredményezõ.
- A Start kodon azonosítása sokkal nehezebb (ATG a metionin kódja is), ld. késõbb.

Kodongyakoriság−eltolódások detektálása

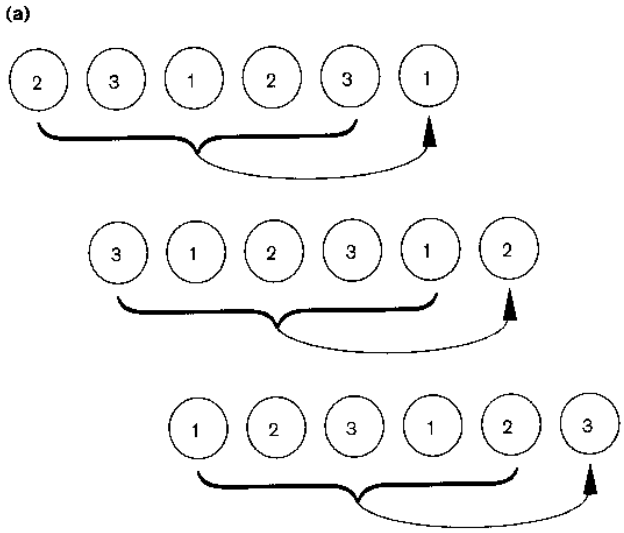
- A nemkódoló és a kódoló régiókban az egyes kodonok (nukleotidhármasok) gyakorisága jellegzetesen eltérõ
- Különféle organizmusok más−más kodonpreferenciát mutatnak. Pl. a szerin hatféle kodonjának gyakoriságai 5 fajban:

Kodon	***E. coli***	***D. melanogaster***	***H. sapiens***	***Z. mays***	***S. cerevisiae***
AGT	3	1	10	3	5
AGC	20	23	34	30	4
TCG	4	17	9	22	1
TCA	2	2	5	4	6
TCT	34	9	13	4	52
TCC	37	48	28	37	33

- Organizmuson belül is változhat! Pl. E. coli−nál három kategória
- Sokféleképpen meg lehet ragadni, legsikeresebb: a **hexamerek** (nukleotidhatosok) elõfordulási gyakorisága
- Pl. GeneMark program valószínûséget számol (kódoló−e a régió) a hexamergyakoriságok alapján, mind a 6 leolvasási keretben.



- Más programok (pl. GenScan) Markov−modellekkel dolgoznak:



Pozíciófüggõ (inhomogén) ötödrendû Markov−modell: minden pozícióban valamely nukleotid valószínûségét az elõtte levõ öt nukleotid milyensége határozza meg, s ez a kerethez viszonyított pozíciótól is függ. A valószínûségértékeket a már ismert gének alapján határozzák meg, új szekvencia valószínûsége a modell alapján kiértékelhetõ.

Funkcionális helyek detektálása

Promoterek, exon−intron határok, transzlációiniciációs helyek, terminációs helyek azonosítása, az ezekre jellemzõ szekvenciamintázatok alapján

Az algoritmusok jóságának két paramétere:

- **Érzékenység:** a valóságos funkcionális helyek mekkora részét jósolja meg helyesen
- **Specificitás:** a megjósolt helyek mekkora része valóságos

Különbözõ megközelítések:

- Konszenzusszekvencia alapján (a leggyakoribb nukleotidok megadása): nem megbízható a variabilitás miatt
- Súlymátrix megadása: minden pozícióra közöljük a 4 nukleotid gyakoriságát, ehhez hasonlítjuk az új szekvenciát. Korlátozott mértékben sikeres.
- Bonyolultabb matematikai modellek (pl. rejtett Markov−modellek). Ígéretes, de nem kiforrott.

Promoterek

- Számos program, az ismert promoterek szekvenciáinak könyvtára alapján
- Pl. TATA−box, "cap" szignál, stb.
- Kevéssé sikeres, mert pl. az emberi gének 30%−ánál nem találhatóak meg a szokásos promoterszignálok
- A sikeresség nem nagyobb, mint a kodongyakoriság−eltolódások alapján

Poliadenilációs szignál

- A gének végén, ált. AATAAA konszenzusszekvenciával, melyet egy bonyolultabb szignál követ
- A gének felénél hiányzik
- Nem megbízható

Transzlációs szignálok: start és stop kodon

- ATG (RNS−ben AUG) a start kodon, de a metionin kódja is
- **Kozak−féle szabályok:** a start kodon sokszor jellegzetes környezetben van, pl.

  CCGCC**AUG**G

  és hasonlók (súlymátrixszal leírható)

- Gyenge az érzékenység (70−80%), de nem genomiális, hanem cDNS esetében jól használható
- Stop kodon: egyértelmû

Exon−intron határok

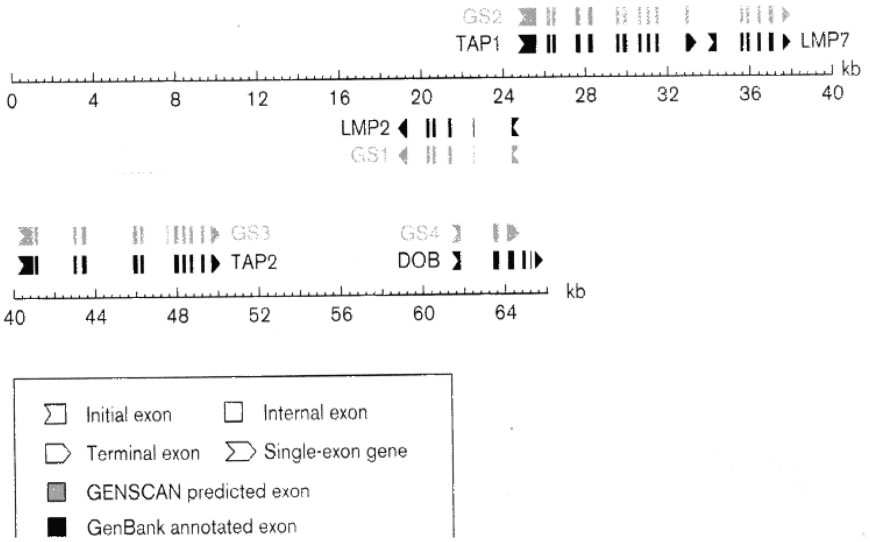
- Az exon−intron határokon (5' vég a donor hely, 3' vég az akceptor) jellegzetes szekvenciák vannak, melyeket a spliceoszóma felismer
- Intronok szinte mindig GT−vel kezdõdnek és AG−vel végzõdnek, ez kevés, de segít leszûkíteni a lehetõségeket
- Bonyolultabb mintázatok súlymátrixa alapján elég hatásos detektálás, de nem tökéletes
- Bonyolultabb matematikai modellek (pl. rejtett Markov) jelentõs, de nem drámai javulást eredményeznek
- Leghatásosabb: együttesen alkalmazni a kodongyakoriság−vizsgálattal (80% fölé emelkedik az érzékenység)

Integrált génelemzés

- Az egyes predikciók hatékonyságát nagyon jelentõsen növeli, ha a különbözõ algoritmusokból származó információt együttesen vesszük figyelembe. Ez vezetett az integrált génelemzõ programok kifejlesztéséhez
- A fenti eljárásokat egymás után alkalmazzák
- Számos ilyen program, szerver van, lásd pl. Gene Prediction Services, Gene Recognition Programs
- 90% körüli érzékenységet, ill. specificitást lehet elérni velük.

Korlátaik

- Csak néhány organizmusra létezik ilyen program
- Néhány kivétellel csak akkor mûködnek jól, ha csak egyetlen gén van a beadott szekvenciában
- Legtöbbször nagyon érzékenyek a szekvenálási hibákra
- Nem képesek kezelni az olyan eseteket, mint pl. alternatív splicing, átfedõ gének, különleges promoterek, stb.
- Ajánlott *mindegyik* programot használni (más−más algoritmusokkal dolgoznak)
- Pl. a Genscan eredményessége egy 66 kilobázis méretû szakaszon:



Fekete jelöli a valóságot, szürke a GenScan általi jóslást (ld. jelmagyarázat). Az LMP2 gén jóslása tökéletes, de a TAP1−et és az LMP7−et a GenScan egybeolvasztotta, a TAP2−be eggyel több exont jósolt, a DOB−nak pedig túl korán véget vetett. Az exonok túlnyomó részét mégis hibátlanul detektálta.

Webhelyek

Ld. pl. Gene Prediction Services, Gene Recognition Programs
