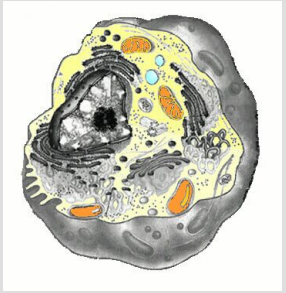
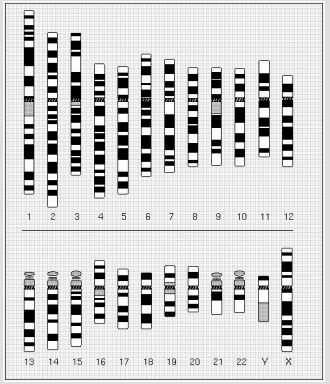
Genomo struktūra

- Genomas





ccgtacgtacgtagagtgctagtctagtcgtagcgccgtagtcgatcgtgtgggtagtagctgatatgatgcgaggtaggggataggat
agcaacagatgagcggatgctgagtgcagtggcatgcgatgtcgatgatagcggtaggtagacttcgcgcataaagctgcgcgagat
gattgcaaagragttagatgagctgatgctagaggtcagtgactgatgatcgatgcatgcatggatgatgcagctgatcgatgtagatgc
aataagtcgatgatcgatgatgatgctagatgatagctagatgtgatcgatggtaggtaggatggtaggtaaattgatagatgctagatc
gtaggtagtagctagatgcagggataaacacacggaggcgagtgatcggtaccgggctgaggtgttagctaatgatgagtacgtatg
aggcaggatgagtgacccgatgaggctagatgcgatggatggatcgatgatcgatgcatggtgatgcgatgctagatgatgtgtgtca
gtaagtaagcgatgcggctgctgagagcgtaggcccgagaggagagatgtaggaggaaggtttgatggtagttgtagatgattgtgta
gttgtagctgatagtgatgatcgtag ……

Genų radimas

- Žmogaus genome yra ~3 milijardai bazių porų ir apie 25000+ baltymus koduojančių baltymų

ccgtacgtacgtagagtgctagtctagtcgtagcgccgtagtcgatcgtgtgggt
agtagctgatatgatgcgaggtaggggataggatagcaacagatgagcggat
gctgagtgcagtggcatgcgatgtcgatgatagcggtaggtagacttcgcgcat
aaagctgcgcgagatgattgcaaagragttagatgagctgatgctagaggtca
gtgactgatgatcgatgcatgcatggatgatgcagctgatcgatgtagatgcaat
aagtcgatgatcgatgatgatgctagatgatagctagatgtgatcgatggtaggt
aggatggtaggtaaattgatagatgctagatcgtaggtagtagctagatgcagg
gataaacacacggaggcgagtgatcggtaccgggctgaggtgttagctaatg
atgagtacgtatgaggcaggatgagtgacccgatgaggctagatgcgatggat
ggatcgatgatcgatgcatggtgatgcgatgctagatgatgtgtgtcagtaagta
agcgatgcggctgctgagagcgtaggcccgagaggagagatgtaggagga
aggtttgatggtagttgtagatgattgtgtagttgtagctgatagtgatgatcgtag
……

Kurgi yra tie genai?

Pagrindinė šablonų atpažinimo idėja

- Kaip vaikai atskiria “katę” nuo “šuns”?
  - Jie buvo apmokyti sakant “A yra šuo”, “B yra katė”, “C yra kitoks šuo”…..
  - Jie išmoksta “išgauti” bendras gyvūnų sąvybes (šablonus) sakant kur yra katė, o kur šuo
  - Tada išgautas sąvybes pritaiko atpažįstant naujus šunis ir kates
- Šablonų atpažinimas paprastai atliekamas:
  - Pateikiant “apmokymo duomenis”, kurie yra pažymėti “teigiamas” arba “neigiamas”, “geras” arba “blogas” ir tt.
  - Išmokstant bendras taisykles kurios atskiria “teigiamą” ir “neigiamą” ir tt.
  - Pritaikant šias taisykles naujoms situacijoms

Genų atpažinimas mokantis

- “Bendrų taisyklių” genų paieškoje mokymasis

ccgtacgtacgtagagtgctagtctagtcgtagcgccgtagtcgatcgtgtgggt
agtagctgatatgatgcgaggtaggggataggatagcaacagatgagcggat
gctgagtgcagtggcatgcgatgtcgatgatagcggtaggtagacttcgcgcat
aaagctgcgcgagatgattgcaaagragttagatgagctgatgctagaggtca
gtgactgatgatcgatgcatgcatggatgatgcagctgatcgatgtagatgcaat
aagtcgatgatcgatgatgatgctagatgatagctagatgtgatcgatggtaggt
aggatggtaggtaaattgatagatgctagatcgtaggtagtagctagatgcagg
gataaacacacggaggcgagtgatcggtaccgggctgaggtgttagctaatg
atgagtacgtatgaggcaggatgagtgacccgatgaggctagatgcgatggat
ggatcgatgatcgatgcatggtgatgcgatgctagatgatgtgtgtcagtaagta
agcgatgcggctgctgagagcgtaggcccgagaggagagatgtaggagga
aggtttgatggtagttgtagatgattgtgtagttgtagctgatagtgatgatcgtag
.....

Per daugelį metų eksperimentiniu keliu buvo atrasta daugybė genų. Taip pat įrodyta, kad kai kurie DNR segmentai yra ne genai.

Genų atpažinimas mokantis

- Taigi mes žinom

genai

ccgtacgtacgtagagtgctagtctagtcgtagcgccgtagtcgatcgtgt

gggtagtagctgatatgatgcgaggtaggggataggatagcaacagatgagc
ggatgctgagtgcagtggcatgcgatgtcgatgatagcggtaggtagacttcgc
gcataaagctgcgcgagatgattgcaaagragttagatgagctgatgctagag

gtcagtgactgatgatcgatgcatgcatg

ne genai

gatgatgcagctgatcgatgtagatgcaataagtcgatgatcgatgatgatgcta
gatgatagctagatgtgatcgatggtaggtaggatggtaggtaaattgatagatg

ctagatcgtaggtagtagctagatgcagggataaacacacggaggc

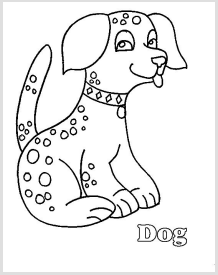
gagtgatcggtaccgggctgaggtgttagctaatgatgagtacgtatgaggcag
gatgagtgacccgatgaggctagatgcgatggatggatcgatgatcgatgcat
ggtgatgcgatgctagatgatgtgtgtcagtaagtaa

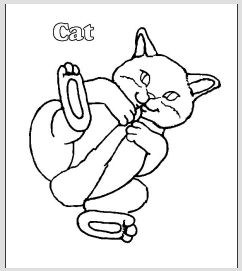
.....

Genų atpažinimas mokantis

- Ar gcgatgcggctgctgagagcgtaggcccgagaggagagat gtaggaggaaggtttgatggtagttgtagatgattgtgtagttgta gctgatagtgatgatcgtag yra genas?

- Prisiminkime “kates” ir “šunis” ….



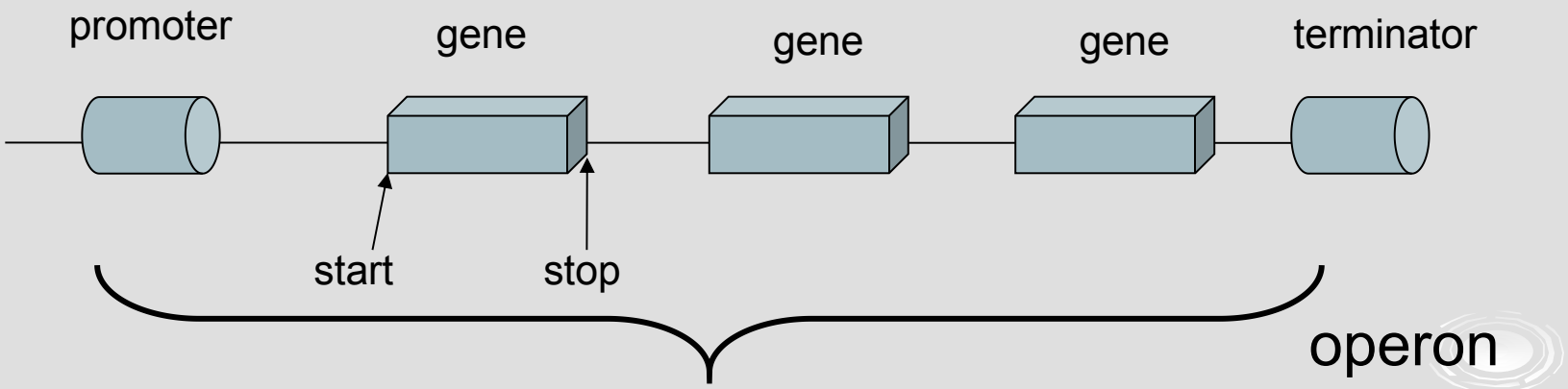


- Šablonai čia yra visai kitokie nei atskiriant katę nuo šuns, ir galbūt labiau “paslėpti” ir sudėtingesni

Taigi pirmiausia reikia išnagrinėti geno sandarą(struktūrą) ….!

Pagrindinės genų struktūros

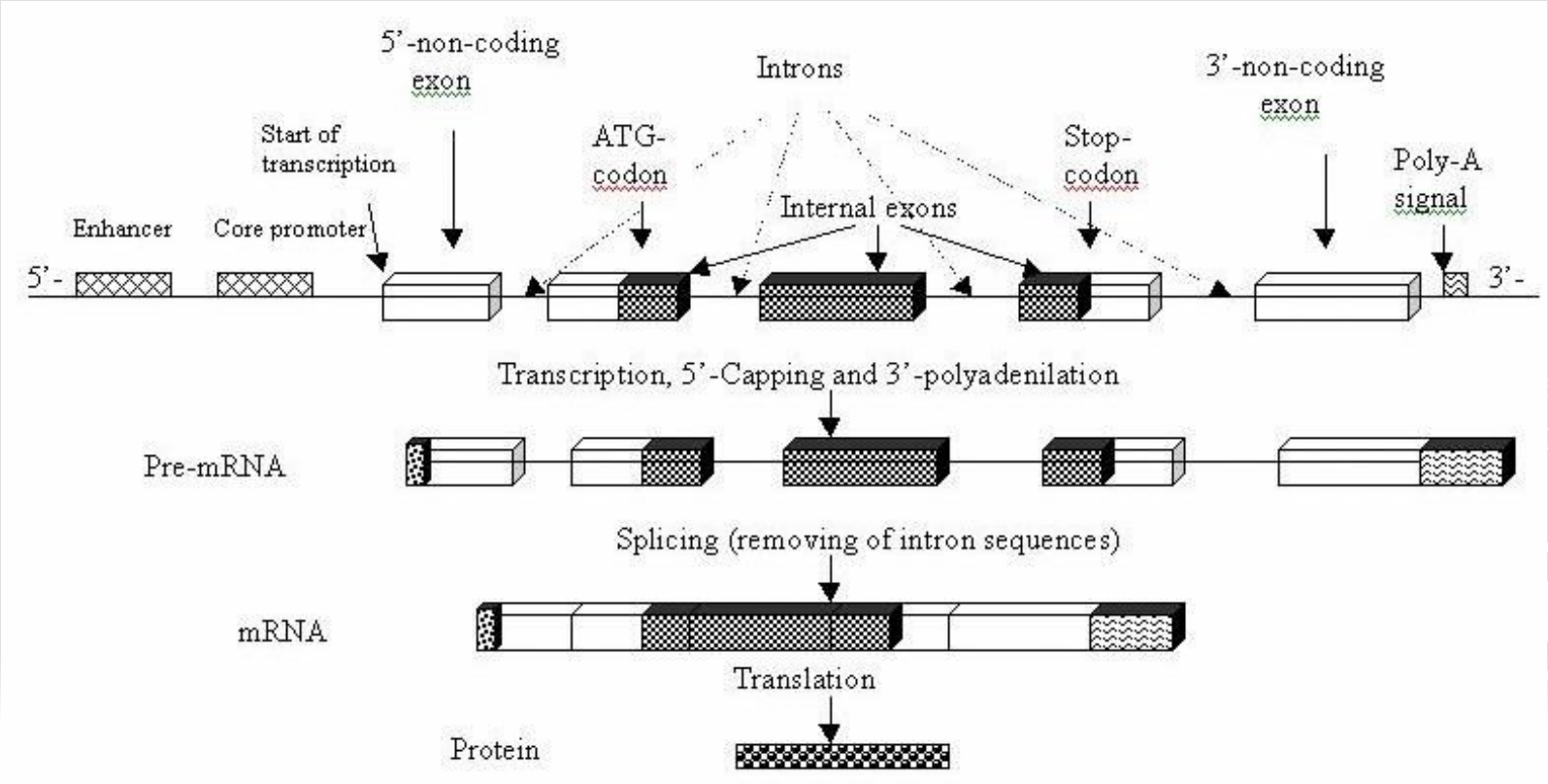
- Prokariotinis genas
  - Koduojantys regionai, nekoduojantys regionai
  - Transliacijos pradžia ir pabaiga



Prokariotinių genų paieška yra žymiai paprastesnė nei eukariotinių genų dėl paprastesnės geno sandaros ir genų tankio genome

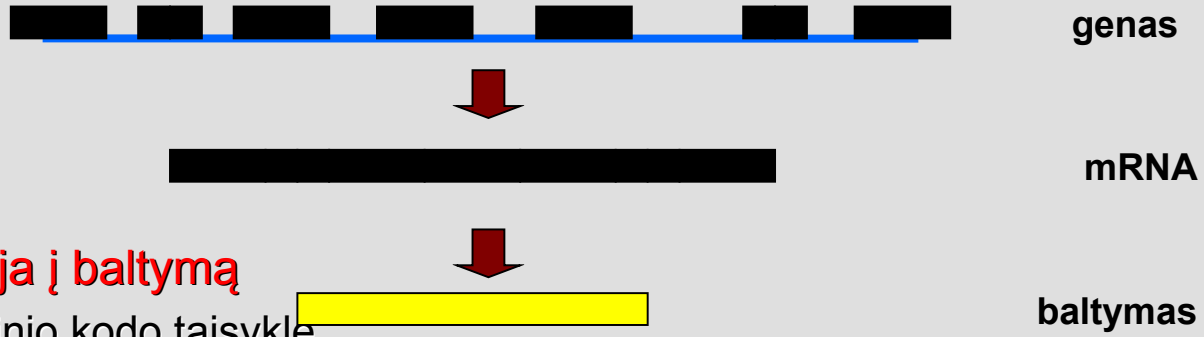
Pagrindinės genų struktūros

- Eukariotiniai genai
  - Egzonai(Exons), intronai(introns),
  - Transliacijos pradžia ir pabaiga, iškirpimo sujungimo vietos



Geno struktūra

- Genų ekspresijos žingsniai
  - Transkripcija:: intronai iškerpami



  - Transliacija į baltymą
    - Genetinio kodo taisyklė

- < 2% žmogaus genomo koduoja baltymus
  - T.y. baltymus koduojantys genai sudaro tik nedidelę genomo dalį

Dvi genų paieškos metodų klasės

- Homologija paremti metodai
  - Genų paieška vykdoma atliekant homogijų paiešką

- *Ab initio* metodai
  - Genų paieška išskiriant specifines sąvybes

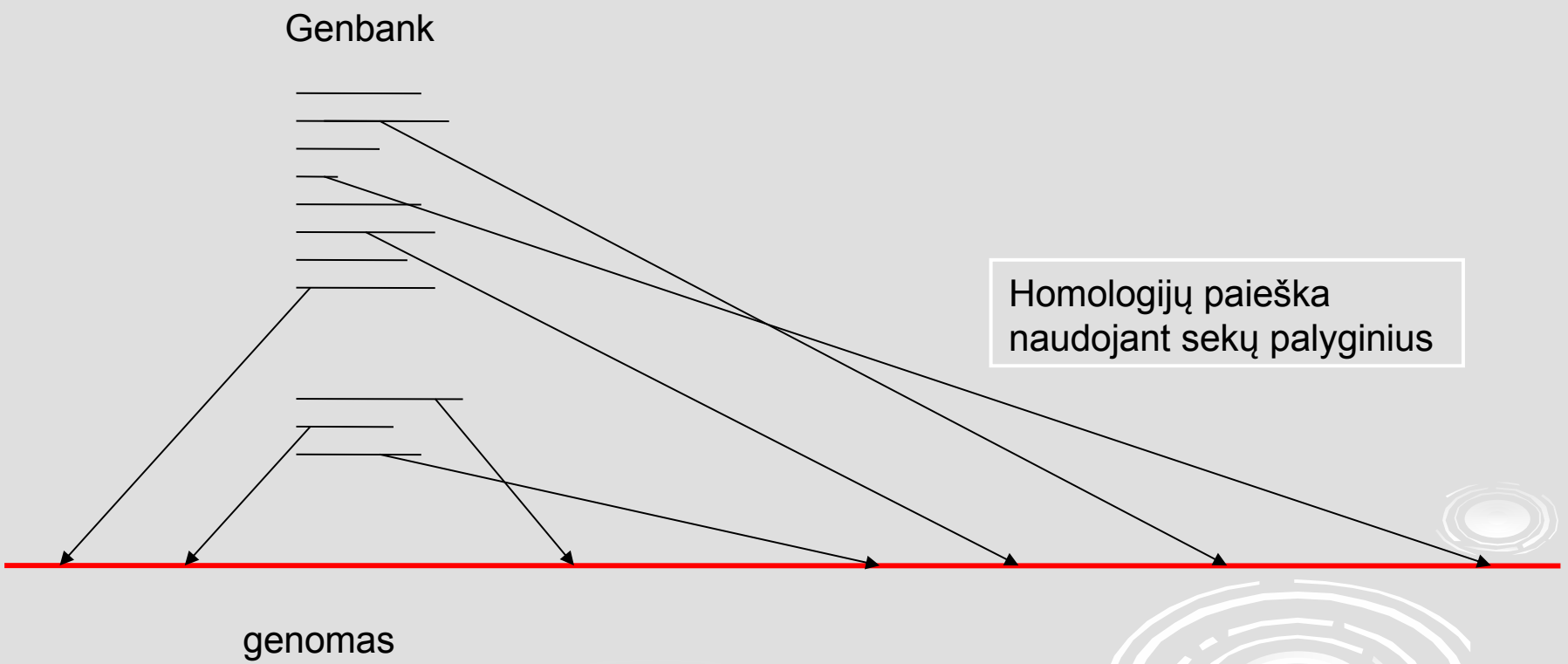
Homologija paremta genų paieška

- Sekos turinčios didelį panašumo laipsnį rodo, kad jos kilusios iš bendro “tėvinio” geno. Tokios sekos vadinamos homologinėmis

- Panašios sekos gali būti atpažintos naudojant sekų palyginio algoritmus
  - BLAST, Smith-Waterman

- *nr* yra visų žinomų genų duomenų bazė; naujame genome suradus sekas panašias į *nr* genus, galima “nuspėti” genus

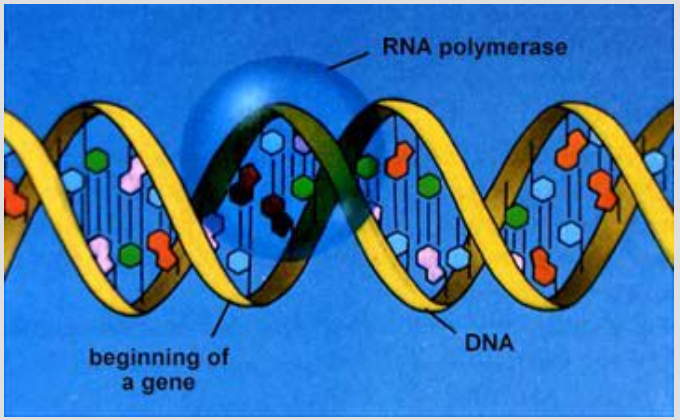
Homologija paremta genų paieška



Genų sąvybės

DNR yra dvigrandė

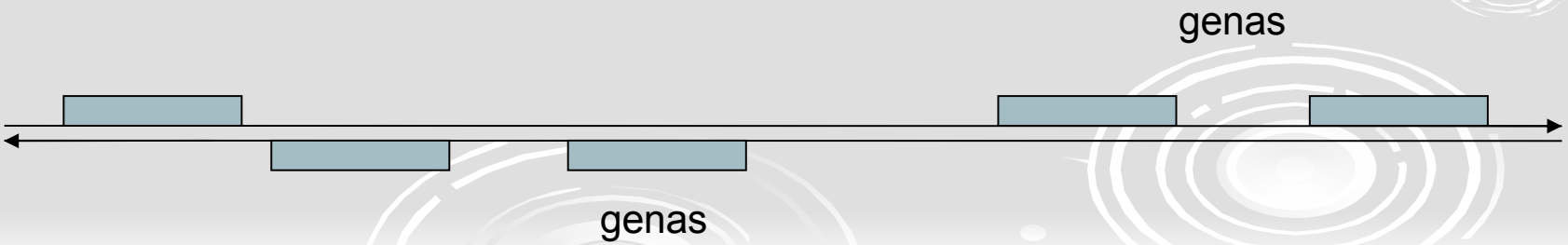
forward strand



…..ACGTTTGA ……

…..TGCAAACT…….

reverse strand



Genų sąvybės

Nukleotidų sekos transliacija į amino rūgščių seką:

Kiekvienas tripletas(kodonas) yra transliuojamas į amino rūgštį

AAATCACGAGAT …….

K H E

Kiekvienas tripletas gali būti nuotransliuotas į amino rūgštį, bet....

Geno struktūra – skaitymo rėmelis (reading frame)

- Skaitymo (arba transliacijos) rėmelis: kiekvienas DNR segmentas turi šešis skaitymo rėmelius

Forward strand: 

Skaitymo rėmelis #0	Skaitymo rėmelis #1	Skaitymo rėmelis #2
ATG	TGG	GGC
GCT	CTT	TTA
TAC	ACG	CGC
GCT	CTT	TTG
TGC	GA.	A..

Reverse strand: 

Skaitymo rėmelis #0	Skaitymo rėmelis #1	Skaitymo rėmelis #2
TCA	CAA	AAG
AGC	GCG	CGT
GTA	TAA	AAG
AGC	GCC	CCA
CAT	AT.	T..

Geno struktūra – atviras skaitymo rėmelis (ORF)

- Atviras skaitymo rėmelis (Open reading frame ,ORF): DNR segmentas su dviem tame pačiame skaitymo rėmelyje esančiais stop kodonais kraštuose ir kuris neturi stop kodono viduryje

Kiekvienas ORF turi fiksuotą skaitymo rėmelį

Kiek genų gali sutalpinti ORF’as?

Atsakymas: vieną, nes ORF’as turi tik vieną stop’ą

Geno struktūra – atviras skaitymo rėmelis (ORF)

- Beveik tiesa: visi ilgi (>300 bp) orf'ai prokariotiniuose genomuose turi genus

- Bet tai ne visada tiesa eukariotiniuose genomuose

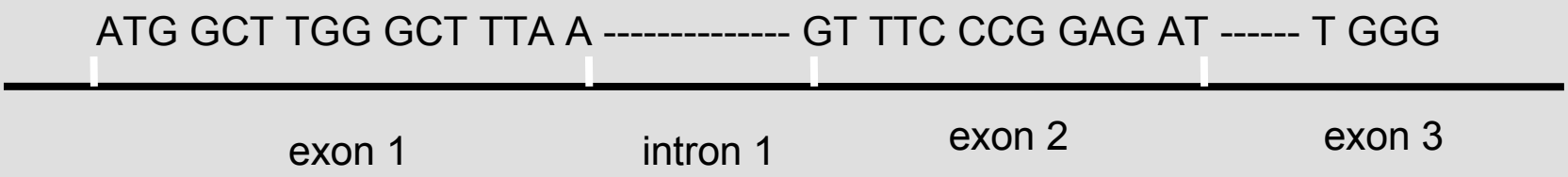
- Koduojantis regionas –
  - Genas prokariotiniuose genomuose
  - Egzonas eukariotiniuose genomuose

Geno struktūra

- Kiekvienas koduojantis regionas (egzonas arba visas genas) turi fiksuotą trasliacijos rėmelį
- Koduojantys regionai visada “sėdi” ORF’o viduje su tuo pačiu skaitymo rėmeliu
- Visi geno egzonai yra toje pačioje grandinėje
- Gretimi geno egzonai gali būti skirtinguose skaitymo rėmeliuose

Geno struktūra – skaitymo rėmelių pastovumas

- O dabar apie “sudėtingesnes” sąvybes
- Gretimi geno egzonai turi būti skaitymo rėmeliais suderinami



egzonas1 [i, j] rėmelyje a ir egzonas2 [m, n] rėmelyje b yra suderinami jei

$$b = (m - j - 1 + a) \bmod 3$$

1 mod 3 = 1
2 mod 3 = 2
3 mod 3 = 0
4 mod 3 = 1
5 mod 3 = 2
.....

Splicing’as!

Kodonų dažnumai

- Koduojačios sekos yra transliuojamos į baltymo sekas
- Nustatyta – dimerų dažnis baltymų sekose yra netolygiai pasiskirstęs

Name	ala	arg	asn	asp	cys	glu	gln	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	9.5	4.1	4.3	5.3	1.2	6	4.8	6.5	2	6.5	11.5	6	2.6	3.7	3.5	6.2	5	1.1	2.7	6.5
arg	7.9	5.5	3.9	5.3	1.1	6	5.5	5.9	2.6	6.5	11.4	5	2.2	4.7	3.6	5.5	4.4	1.4	4	6.6
asn	9.6	4.9	4.2	4.9	1	5.3	5.6	7.4	2.3	6	10	4.9	2	3.5	5.1	6.1	5.5	1.5	3.1	6.1
asp	9.3	4	4.7	5.1	1	6.7	2.9	7	1.8	7.1	9.6	6.3	2.3	4.3	3.9	5.9	5.1	1.6	3.6	6.6
cys	8.4	4.8	3.3	5.4	1.7	5.6	5.2	8.1	4.3	5.4	10.2	3.8	1.8	4.1	4.5	6.3	4.3	1.6	3.4	6.8
glu	9.4	5.8	3.6	4.5	0.8	4.9	7	5.8	2.6	5.9	12.7	5	2.4	4	3.5	5.4	5	1.1	2.8	6.8
gln	10.3	4.9	3	4.4	0.9	4.5	6.8	7	2.7	5.5	12.8	4.1	2	3.9	3.8	5.8	5.3	1.4	3	6.9
gly	8.1	4.8	3.9	5.1	1.2	6	4.6	6.4	2.4	6.8	10.5	5.8	2.7	4.8	2.4	5.8	5.1	1.4	3.7	7.5
his	7.3	4.7	4	4.8	1.5	4.9	5.6	6.9	3	6.2	10.8	4.8	1.6	5	5.2	6.8	4.9	1.7	4.2	5.1
ile	11	4.7	4.9	6.5	1.1	6.9	3.6	7.2	2.1	5.3	8.6	5.3	1.8	3.2	4.2	7	5.6	0.9	2.9	6.1
leu	10.4	4.2	4.3	5.2	1.1	5.2	3.7	6.8	2	5.6	10.6	5.3	2.3	3.8	4.5	7.4	6.2	1	2.6	6.6
lys	10.6	5.2	3.8	5.2	0.5	5.3	5.9	6.6	2.6	5.2	11.3	4.7	1.9	2.8	4.6	6	5.5	1.2	2.6	7.6
met	10.8	4.8	3.8	4.6	0.7	4.6	4.9	7	1.7	4.7	11.4	5.2	2.8	3.3	5.1	7.4	6.3	0.9	2	6.8
phe	9.6	3.7	5.2	6.5	1.2	6.4	2.7	7.9	1.9	6.7	7.4	5	2.5	3.9	3.6	8	5.8	1.3	3.3	6.3
pro	8.4	3.6	4.6	5.4	0.7	7.6	5.2	5.4	2.3	6.1	11.2	5.5	2.4	4.2	2.8	6.5	5.4	1.4	2.9	7.5
ser	9.1	4.6	3.7	5	1	5.4	5.2	7.2	2.6	6	11.6	4.5	2.2	4.1	4.1	6.5	5	1.2	3.2	6.8
thr	9.1	4.2	3.7	5.6	0.9	5.7	5.7	7.5	2.2	5.5	12	4.2	2	3.5	5.5	6.2	5.3	1.1	2.6	6.7
trp	7.1	6.3	3.2	4.8	1.3	3.9	8.5	6.6	3.6	5	14.2	3.2	2.4	4.6	3.9	5.8	4.3	1.3	3	6.1
tyr	7.9	6.5	3.6	4.9	1.2	4.5	7	7.1	2.6	5	11.7	4	1.6	4.7	4.9	6.4	4.6	1.5	3.4	5.7
val	9.6	4.1	4.4	5.9	1	6.2	3.4	6.4	1.8	6.5	10.2	5.2	2.5	3.7	3.8	7.2	6.1	1.1	2.7	7.1

Vidutinis dažnis 5%

Kai kurios amino rūgštys linkusios būti viena šalia kitos

Kai kurios amino rūgštys linkusios **ne**būti viena šalia kitos

shewanella

Dikodonų dažnumai

- Netolygiu dimerų dažnių pasiskirstymu paremtas daugelio genų paieškos programų veikimas!

- Pagrindinė genų paieškos idėja – jei dimeras turi mažesnį dažnį nei vidutinis, tai reiškia kad baltymas nelinkęs “turėti” tokių dimerų; kitu atveju baltymas linkęs “turėti” tokius dimerus

Taigi jei matome kad dikodonas koduoja dimerą, galime “statyti” kad tai nėra koduojantis regionas!

Dikodonų dažnumai

- Taigi jei matome daug tokių dikodonų DNR segmente, galime teigti, kad tai nekoduojantis regionas!

Tai ir yra pagrindinė genų paieškos algoritmų idėja!

Dikodonų dažnumai

- Dikodonų dažnumai yra nuo genomo priklausomi

Name	ala	arg	asn	asp	cys	glu	gln	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	9.5	4.1	4.3	5.3	1.2	6	4.8	6.5	2	6.5	11.5	6	2.6	3.7	3.5	6.2	5	1.1	2.7	6.5
arg	7.9	5.5	3.9	5.3	1.1	6	5.5	5.9	2.6	6.5	11.4	5	2.2	4.7	3.6	5.5	4.4	1.4	4	6.6
asn	9.6	4.9	4.2	4.9	1	5.3	5.6	7.4	2.3	6	10	4.9	2	3.5	5.1	6.1	5.5	1.5	3.1	6.1
asp	9.3	4	4.7	5.1	1	6.7	2.9	7	1.8	7.1	9.6	6.3	2.3	4.3	3.9	5.9	5.1	1.6	3.6	6.6
cys	8.4	4.8	3.3	5.4	1.7	5.6	5.2	8.1	4.3	5.4	10.2	3.8	1.8	4.1	4.5	6.3	4.3	1.6	3.4	6.8
glu	9.4	5.8	3.6	4.5	0.8	4.9	7	5.8	2.6	5.9	12.7	5	2.4	4	3.5	5.4	5	1.1	2.8	6.8
gln	10.3	4.9	3	4.4	0.9	4.5	6.8	7	2.7	5.5	12.8	4.1	2	3.9	3.8	5.8	5.3	1.4	3	6.9
gly	8.1	4.8	3.9	5.1	1.2	6	4.6	6.4	2.4	6.8	10.5	5.8	2.7	4.8	2.4	5.8	5.1	1.4	3.7	7.5
his	7.3	4.7	4	4.8	1.5	4.9	5.6	6.9	3	6.2	10.8	4.8	1.6	5	5.2	6.8	4.9	1.7	4.2	5.1
ile	11	4.7	4.9	6.5	1.1	6.9	3.6	7.2	2.1	5.3	8.6	5.3	1.8	3.2	4.2	7	5.6	0.9	2.9	6.1
leu	10.4	4.2	4.3	5.2	1.1	5.2	3.7	6.8	2	5.6	10.6	5.3	2.3	3.8	4.5	7.4	6.2	1	2.6	6.6
lys	10.6	5.2	3.8	5.2	0.5	5.3	5.9	6.6	2.6	5.2	11.3	4.7	1.9	2.8	4.6	6	5.5	1.2	2.6	7.6
met	10.8	4.8	3.8	4.6	0.7	4.6	4.9	7	1.7	4.7	11.4	5.2	2.8	3.3	5.1	7.4	6.3	0.9	2	6.8
phe	9.6	3.7	5.2	6.5	1.2	6.4	2.7	7.9	1.9	6.7	7.4	5	2.5	3.9	3.6	8	5.8	1.3	3.3	6.3
pro	8.4	3.6	4.6	5.4	0.7	7.6	5.2	5.4	2.3	6.1	11.2	5.5	2.4	4.2	2.8	6.5	5.4	1.4	2.9	7.5
ser	9.1	4.6	3.7	5	1	5.4	5.2	7.2	2.6	6	11.6	4.5	2.2	4.1	4.1	6.5	5	1.2	3.2	6.8
thr	9.1	4.2	3.7	5.6	0.9	5.7	5.7	7.5	2.2	5.5	12	4.2	2	3.5	5.5	6.2	5.3	1.1	2.6	6.7
trp	7.1	6.3	3.2	4.8	1.3	3.9	8.5	6.6	3.6	5	14.2	3.2	2.4	4.6	3.9	5.8	4.3	1.3	3	6.1
tyr	7.9	6.5	3.6	4.9	1.2	4.5	7	7.1	2.6	5	11.7	4	1.6	4.7	4.9	6.4	4.6	1.5	3.4	5.7
val	9.6	4.1	4.4	5.9	1	6.2	3.4	6.4	1.8	6.5	10.2	5.2	2.5	3.7	3.8	7.2	6.1	1.1	2.7	7.1

shewanella

bovine

Name	ala	arg	asn	asp	cys	glu	gln	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	11.4	5.9	3.1	4.5	1.9	5.8	3.6	7.7	1.9	4.3	9.7	4.3	2.1	3.7	6.4	6.4	5.6	1.1	2.6	6.8
arg	8.5	7.7	4	4.6	2.3	5.9	3.8	7.6	2.5	4.4	9.2	5	1.7	4	5.3	6.3	5	1.5	3.4	6.5
asn	6.3	4.9	4.9	4.4	2.1	5.3	4.1	6.9	2.2	5.6	9.7	5.4	2.1	4.1	5.9	7.3	5.3	1.9	4.6	6.2
asp	7.4	4.9	3.5	5.4	2.4	6.6	3.4	7.4	2.1	5.4	9.5	4.7	2	4.4	5.4	6.8	5.7	1.6	4	6.4
cys	6.9	5.9	4	5.4	2.7	5.6	4.9	7.1	3	4.4	8.8	5.4	1.6	3.5	6.8	7.4	5.7	1.4	2.7	5.7
glu	7.8	5.3	4.3	6.4	1.9	9.7	3.7	6.8	2	5.1	8.2	6.2	2.2	3.3	4.8	5.3	5.4	1.2	3.2	6.2
gln	7.9	5.6	4.2	5	2	6.6	5.1	6.9	2.1	4.7	9.3	5.7	2	3.3	5.9	5.7	6.1	1.6	3.3	6.2
gly	7.9	5.8	3.9	5	1.9	6.2	3.5	8	1.8	4.7	8.7	5.2	1.7	3.7	6.9	7.4	5.8	1.4	3.2	6.2
his	6	5.8	4.3	3.5	2.9	5.1	4.1	6.3	3.2	4.5	10.6	4.8	1.6	4.5	6.7	6.6	6.1	1.7	3.9	6.9
ile	6.2	4.9	4.9	4.7	2.4	5.3	4.6	5.8	2.2	6	9.9	5.3	2.1	4.1	5.3	7.7	6.9	1.2	3.7	6
leu	7.7	5.6	4.1	4.7	2.1	5.8	4.5	6.8	2.1	4.6	11	5.4	1.9	3.7	5.7	7	5.5	1.2	3.1	6.4
lys	6.3	5.2	4.8	5.2	2.1	7.2	3.7	6.7	2.2	6	8.5	7.5	2	3.5	4.8	6.1	5.8	1.6	3.5	6.3
met	9.3	5.3	4.1	5.9	1.6	6.1	3.5	6.4	1.6	4.1	9.6	6.6	2.6	4	5.1	6.9	5.5	1	3.2	6.6
phe	6	5.4	4.5	5.2	2.5	5.5	4.1	6.5	2.3	5.3	10.2	5.2	1.8	4.1	5.3	7.8	5.8	1.4	3.9	6.2
pro	8.5	5.4	3.1	5.1	1.9	6.7	3.9	9.5	1.9	4.3	7.7	4.3	1.7	3.3	8.7	6.9	5.7	1.4	2.8	6.4
ser	6.7	5.4	3.8	4.9	2.3	5.4	4	7.9	2.1	4.5	9.5	5.2	1.8	4	5.7	8.6	6.2	1.4	3	6.4
thr	7.5	4.6	3.7	5	2.6	5.7	3.8	6.8	2	5.2	9.7	4.4	1.8	3.9	6	7.2	7.3	1.5	3.5	6.9
trp	7.1	5.2	4.9	5.5	2.3	5.4	4.3	5.8	2.2	5.6	9.5	6.6	2.1	3.8	4.1	6.4	5.9	1.7	3.7	6.8
tyr	5.8	5.7	5	5.1	2.3	5.7	4.1	6.2	2.4	5	8.6	5.6	1.9	5	4.8	6.7	6.3	1.5	4.8	6.5
val	7.6	5	4.4	5.2	2.4	5.7	3.7	6.3	1.9	5	9.3	5.1	2.1	4.1	5.5	6.9	6.6	1.1	3.6	7.4

Dikodonų dažnumai

- Santykinis dikodonų dažnumas koduojančiuose palyginus su nekoduojančiais regionais
  - Dikodono X (pvz AAA AAA) dažnumas koduojančiame regione, santykis tarp X pasitaikymo skaičiaus ir visų dikodonų skaičiaus
  - Dikodono X dažnumas nekoduojančiame regione, santykis tarp X pasitaikymo skaičiaus ir visų dikodonų skaičiaus

Žmogaus genome dikodono “AAA AAA” dažnis koduojančiuose regionuose yra ~1%, o nekoduojančiuose ~5%

Klausimas: jei jūs matote regioną su daug “AAA AAA”, atspėkite, ar tai koduojantis regionas?

Pagrindinė genų paieškos idėja

- Dauguma dikodonų yra “pasislinkę” į koduojančius arba nekoduojančius regionus. Tik dalis jų yra neutralūs

- Koduojančių regionų identifikavimo pagrindas

Regionas turintys daug dikodonų linkusių būti koduojančiuose regionuose greičiausiai yra koduojantis; ir priešingai - nekoduojantis

- Dikodonų dažniai yra bazinis signalas atpažįstant koduojančius regionus. Visos genų paieškos programos naudoja šią informaciją.

Kompiuterinis modelis genų paieškai

- Polinkio modelis:
  - Kiekvienam dikodonui X (pvz AAA AAA) apskaičiuoti jų dažnius koduojančiuose ir nekoduojančiuose regionuose, FC(X), FN(X)
  - apskaičiuoti X's "polinkio reikšmę P(X) = log (FC(X)/FN(X))

- Sąvybės:
  - P(X) yra 0 , jei X turi tokį patį dažnį koduojančiuose ir nekoduojančiuose regionuose
  - P(X) yra teigiamas, jei X dažniau pasitaiko koduojančiuose regionuose; kuo didesnis skirtumas , tuo teigiamesnė reikšmė
  - P(X) yra neigiamas, jei X dažniau sutinkamas nekoduojančiuose regionuose; kuo didesnis skirtumas, tuo neigiamesė reikšmė

Kompiuterinis modelis genų paieškai

- Pavyzdys

AAA ATT, AAA GAC, AAA TAG turi tokius dažnumus

FC(AAA ATT) = 1.4%, FN(AAA ATT) = 5.2%
FC(AAA GAC) = 1.9%, FN(AAA GAC) = 4.8%
FC(AAA TAG) = 0.0%, FN(AAA TAG) = 6.3%

Tada

P(AAA ATT) = log (1.4/5.2) = -0.57
P(AAA GAC) = log (1.9/4.8) = -0.40
P(AAA TAG) = - begalybė (STOP kodonai traktuojami skirtingai)

Regionas sudarytas iš šių dikodonų greičiausiai nekoduojantis

- Regiono kodavimo polinkis

Apskaičiuojamas visų regiono dikodonų kodavimo “polinkiai” ir susumuojami;

Jei suminė reikšmė yra teigiama, spėjame kad regionas koduojantis, priešingu atveju - nekoduojantis

Kompiuterinis genų radimas

- Koduojančio regiono paieškos procedūra

Procedūra:

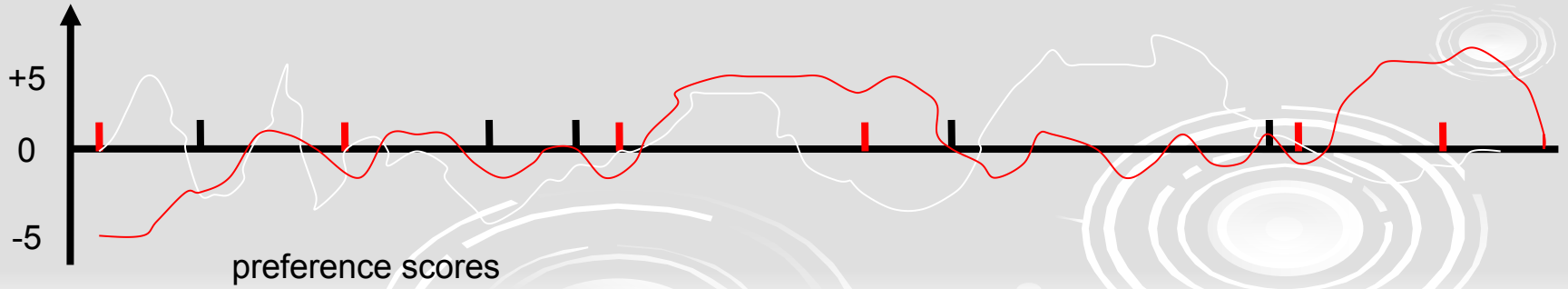
Rasti visus DNR segmento ORF'us;

Kiekvienam ORF'ui

sliktis per ORF'ą su 10 bp inkrementu

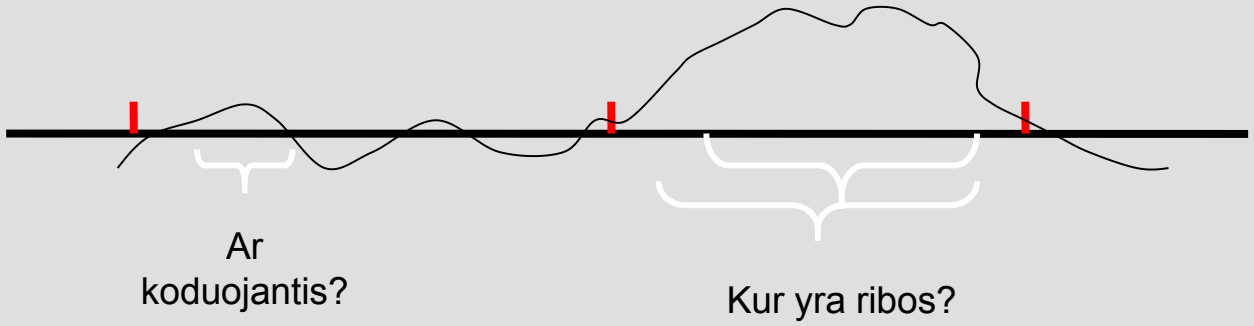
apskaičiuoti "polinkio" reikšmę tame pačiame skaitymo rėmelyje, naudojant 60 bp langą. Priskirti apskaičiuotą reikšmę lango viduriui

pavyzdys (viršutinė grandinė, vienas skaitymo rėmelis)



Kompiuterinis genų radimas

- Išsiaiškinti: koduojantis ar nekoduojantis ir kur yra ribos



- Reikalingi apmokymo duomenys su koduojančiais ir nekoduojančiais regionais
  - Pasirinkti ribines reikšmes kad kuo daugiau koduojančių regionų tilptų ir tuo pačiu atmestų nekoduojančius regionus

Kuris geriausias?

Jei threshold = 0.2, aprėpsime 90% koduojančių regionų ir 10% nekoduojančių regionų

If threshold = 0.4, aprėpsime 70% koduojančių regionų ir 6% nekoduojančių regionų

If threshold = 0.5, aprėpsime 60% koduojančių regionų ir 2% nekoduojančių regionų

Kompiuterinis genų radimas

- Kodėl dikodonas (6meras)?

Kodonais (3merais) –paremti modeliai yra ne tokie informatyvūs kaip dikodonais paremti modeliai

Trikodonais (9merais)-paremti modeliai nenaudojami, nes jiems reikia daug informacijos

Atlikti tyrimai su 7-merais ir 8-merais, jie gali duoti geresnius nuspėjimus

Yra

4*4*4 = 64 kodonų

4*4*4*4*4*4 = 4,096 dikodonų

4*4*4*4*4*4*4*4*4= 262,144 trikodonų

Kad statistika būtų patikima, reikia bent ~15 X-mero pasitaikymų. Taigi trikodonams reikėtų15*262144 = 3932160 koduojančių bazių apmokymo duomenyse, o daugumai genomų tiek duomenų neturime

Kompiuterinis genų radimas

- “Markov chain” modeliai
- “Hidden Markov” modeliai
- Ir sudėtingesni modeliai

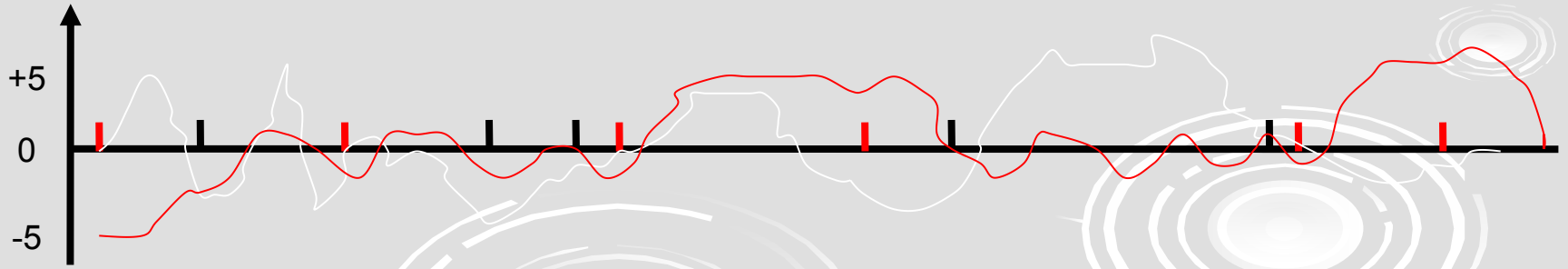
Duomenų surinkimas

- Iš kur galime gauti duomenis apie dikodonų dažnumus?

- Genbank http://www.ncbi.nlm.nih.gov/Genbank/index.html
  - Koduojantys ir nekoduojantys regionai

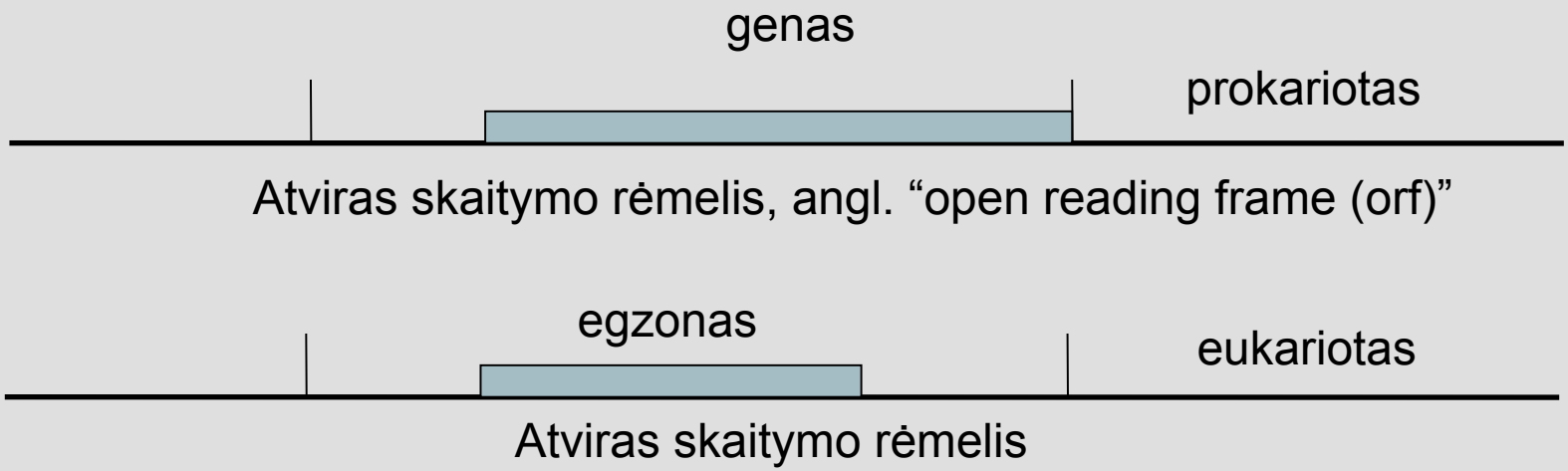
Paprasto genų paieškos variklio sukūrimas

- Susirinkti informaciją apie koduojančius ir nekoduojančius regionus
- Susikurti polinkio modelius (arba sudėtingesnius modelius)
- Kiekvienam skaitymo rėmeliui nuskanuoti genominę seką apskaičiuojant kiekvieno segmento reikšmes (pvz. 60bp)

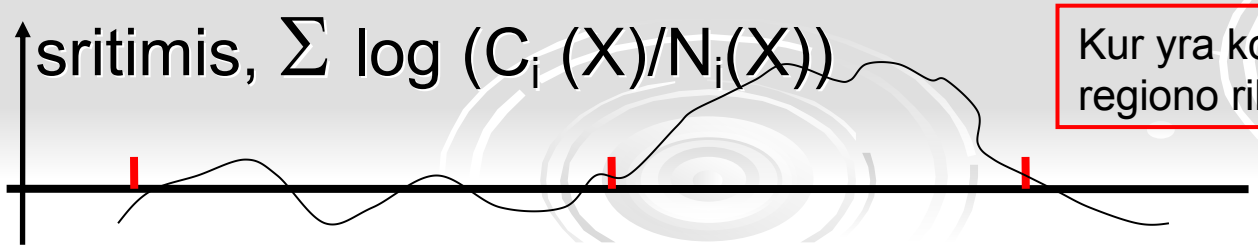


Koduojantys regionai

- Koduojantys regionai



- Kodavimo potencialas – heksamero dažnumas koduojančiose srityse palyginus su nekoduojančiomis sritimis, $\Sigma \log (C_i(X)/N_i(X))$



Kur yra koduojančio regiono ribos?

Ribos signalų nustatymas

- Žinant kokios būna koduojančių regionų ribos, šių regionų identifikacija tampa tikslesnė
- Galimos egzono ribos:

{ transliacijos pradžia, akceptorinė vieta }

{ transliacijos pabaiga, donorinė vieta }

- Transliacijos pradžia
  - Rėmelyje esanti ATG
- Iškirpimo-sujungimo vietos:
  - Donorinė vieta: koduojantis regionas | GT
  - Akceptorinė vieta: YAG | koduojantis regionas. Y žymi C arba T
- Stop kodonas: TAA|TAG|TGA

Transliacijos pradžios radimas

- Transliacijos pradžia: ATG

- Kaip surasti transliacijos pradžią?

GCCATGGCGA …..

ACGATGCTGT ….

GACATGGTAC …

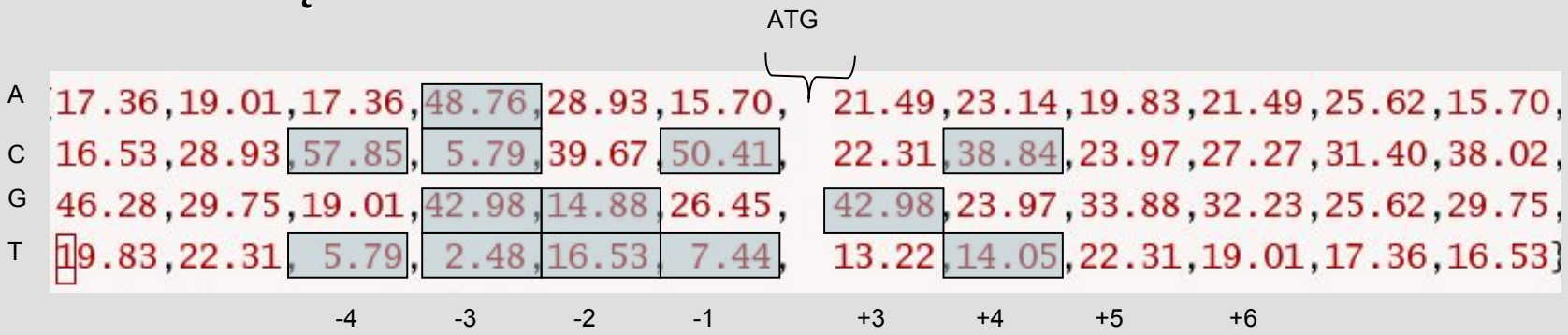
AGGATGGGCT …

GCGATGTGGC …

- Surinkti eksperimentiškai patvirtintų transliacijos pradžių sekas su supančiais(flankuojančiais) regionais ir juos sulyginti

Transliacijos pradžios radimas

- Kai kurie nukleotidai linkę būti aplink startą “ATG”, o kai kurie linkę nebūti



- “Tendencingas” nukleotidų pasiskirstymas neša informaciją. Tai ir yra transliacijos pradžios nustatymo pagrindas

- Kylas klausimas: kuri iš sekų labiau linkusi būti transliacijos pradžia? CACC ATG GC

  TCGA ATG TT

Transliacijos pradžios radimas

- Matematinis modelis: Fi (X): X (A, C, G, T) dažnis padėtyje *i*
- Susumuoti visą eilutę Σ log (Fi (X)/0.25)

CACC ATG GC

log (58/25) + log (49/25) + log (40/25) + log (50/25) + log (43/25) + log (49/25) =

0.37 + 0.29 + 0.20 + 0.30 + 0.24 + 0.29

= 1.69

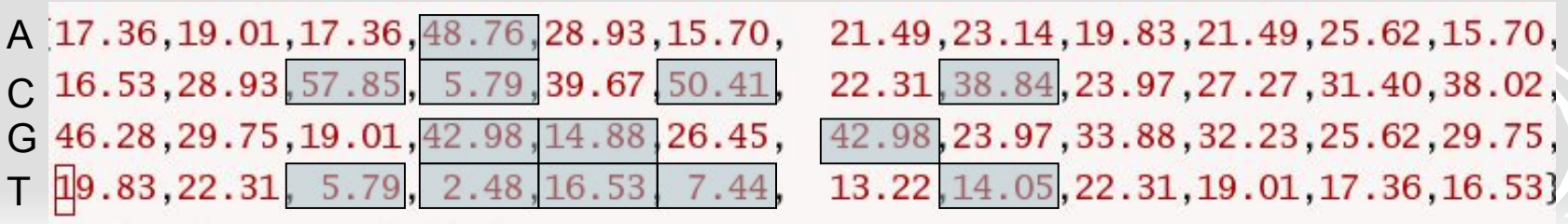
TCGA ATG TT

log (6/25) + log (6/25) + log (15/25) + log (7/25) + log (13/25) + log (14/25) =

-(0.62 + 0.62 + 0.22 + 0.55 + 0.28 + 0.25)

= -2.54

Intuicija neapgavo..



Transliacijos pradžios radimas

- Sukuriamas matematinis modelis, kuris paremtas transliacijos pradžios seka
- Kiekvienai “kandidatinei” sekai pritaikyti modelį ir apskaičiuoti įvertį

- Jei įvertis didesnis nei nulis, bandome spėti kad tai tikėtina transliacijos pradžia; kuo didenė įverčio reikšmė, tuo didesnė spėjimo tikimybė

Iškirpimo-sujungimo vietų radimas

- Iškirpimo-sujungimo vietos (toliau jungtys):
  - Donorinė vieta: koduojantis regionas | GT
  - Akceptorinė vieta: YAG | koduojantis regionas

- Kaip ir transliacijos pradžios, jungtis supančiose sekose nukleotidai pasiskirstę tendencingai
- Šios pasiskirstymo tendencijos yra pagrindas jungtims surasti

Jungčių radimas

- Akceptorių supančios sekos nukleotidų pasiskirstymas

$Y_{75}Y_{72}Y_{78}Y_{79}Y_{77}Y_{80}Y_{66}Y_{78}Y_{85}Y_{84}NC_{68}AG|G_{63}$

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1
A	11.1	12.7	3.2	4.8	12.7	8.7	16.7	16.7	12.7	9.5	26.2	6.3	100	0.0	21.4
C	36.5	30.9	19.1	23.0	34.9	39.7	34.9	40.5	40.5	36.5	33.3	68.2	0.0	0.0	7.9
G	9.5	10.3	15.1	12.7	8.7	9.5	16.7	4.8	2.4	6.3	13.5	0.0	0.0	100	62.7
U	38.9	41.3	58.7	55.6	42.1	40.5	30.9	37.3	44.4	47.6	27.0	25.4	0.0	0.0	7.9

Daug padėčių neša daug informacijos

Informacijos kiekis: Σ F (X) log (F (X)/0.25)

Akceptorinių vietų radimas

- Matematinis modelis: Fi (X): X (A, C, G, T) dažnumas padėtyje i
- Įvertinti segmento akceptorinės vietos galimybę Σ log (Fi (X)/0.25)

- Kiekvienai kandidatinei akceptorinei sekai pritaikomas modelis ir apskaičiuojamas įvertis

- Jei įvertis didesnis nei nulis, bandome spėti kad tai tikėtina akceptorinė vieta; kuo didenė įverčio reikšmė, tuo didesnė spėjimo tikimybė

Donorinių vietų radimas

- Nukleotidų pasiskirstymas supančiose sekose

	-3	-2	-1	1	2	3	4	5	6
A	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
C	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
G	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
U	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

- Matematinis modelis: Fi (X): X (A, C, G, T) dažnumas padėtyje I
- Įvertinti segmento donorinės vietos galimybę Σ log (Fi (X)/0.25)

Donorinių vietų radimas

- Kiekvienai kandidatinei donorinei sekai pritaikomas modelis ir apskaičiuojamas įvertis

- Jei įvertis didesnis nei nulis, bandome spėti kad tai tikėtina donorinė vieta; kuo didenė įverčio reikšmė, tuo didesnė spėjimo tikimybė

Donorų/akceptorių radimas

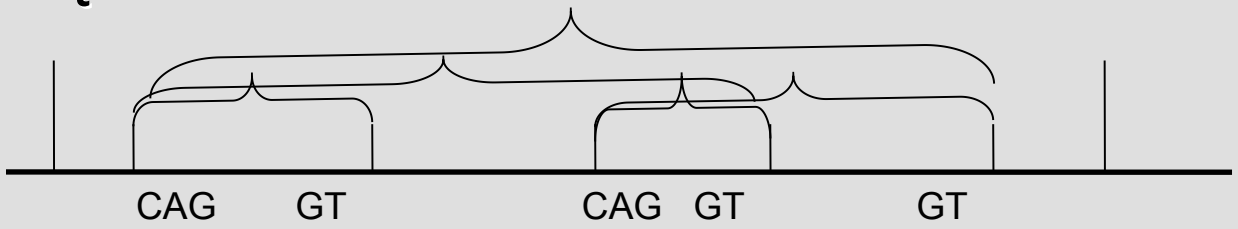
- Nuo padėties priklausomų svorių matrica

	-3	-2	-1	1	2	3	4	5	6
A	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
C	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
G	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
U	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

- Sukurti nuo padėties priklausomų svorių matricos modelį:
  - Surinkti žinomas {donoro, akceptoriaus} sekas ir jas sulyginti
  - Apskaičiuoti nukleotido dažnumą kiekvienoje padėtyje
- Yra ir “gudresnių” modelių norint surinkti daugiau informacijos

Egzonų radimas

- Kiekvienam ORF'ui reikia surasti kandidatines donorų ir akceptorių vietas ieškant GT ir YAG motyvų



- Įvertinti kiekvieną donorą ir akceptorių naudojantis nuo padėties priklausoma svorių matrica
- Surasti visas poras (akceptorius, donoras) kurios turi didesnes nei ribines reikšmes
- Įvertinti segemento (donoras, akceptorius) kodavimo potencialą naudojantis heksamerų modeliu

Egzonų radimas

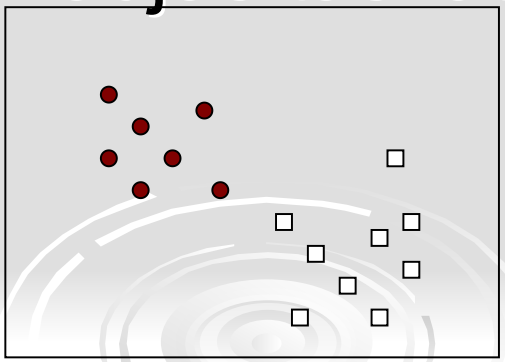
- Taigi kiekvienam segmentui [akceptorius, donoras] turime tris įverčius (kodavimo potencialas, donoro įvertis, akceptoriaus įvertis)
- Variantai
  - Visos reikšmės didelės – greičiausiai tai egzonas
  - Visos reikšmės žemos – greičiausiai tai nėra egzonas
  - Visos reikšmės vidutinės – kas tada?
  - Kai kurios reikšmės didelės, o kai kurios – žemos ?
- Taigi kokios yra egzonų radimo taisyklės

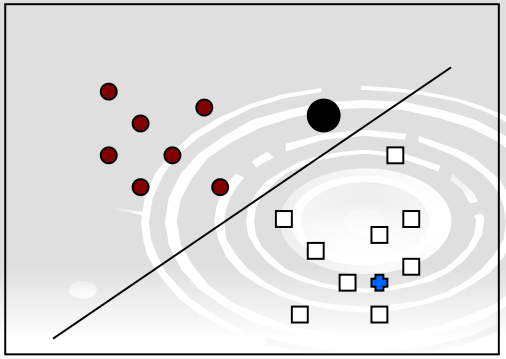
Egzonų radimas

- Mokymasis atskirti egzonus nuo neegzonų
  - Surinkti duomenų apie egzonus ir neegzonus
  - Įvertinti juos
  - Pažymėti buvimo vietą
  - “Išvesti” liniją skiriančią egzonus nuo neegzonų
- “Spėjimą” vykdant atsižvelgti kurioje linijos pusėje yra naujas taškas

koduojantis:

Nekod.:

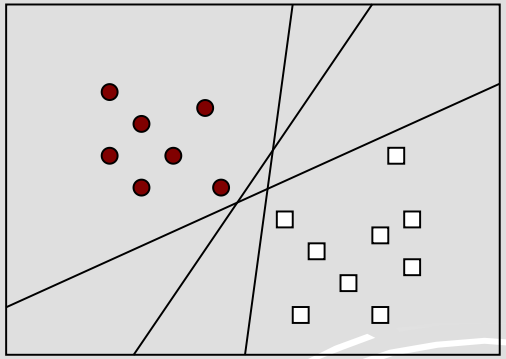




Egzonų radimas

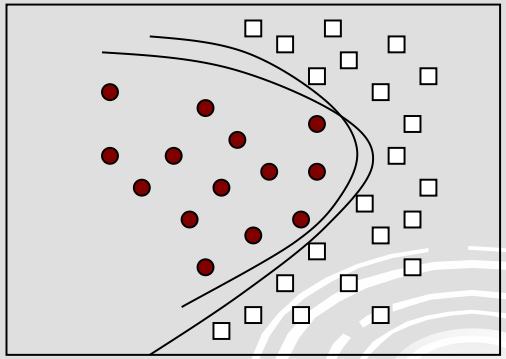
- Atskiriančios linijos nustatymas:
  - Kai kuriais atvejais dvi klasės gali būti atskirtos tiese
  - Bet gali būti kad atskirti pavyks tik su netiesine funkcija

Linijinis diskriminatorius



$Y = a X + b$

Kvadratinis diskriminatorius

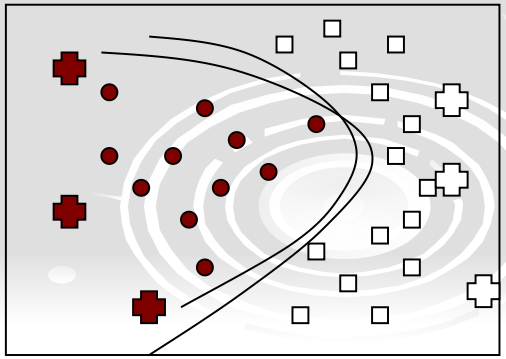


$Y = a X^2 + b X + c$

Diskriminantinė analizė

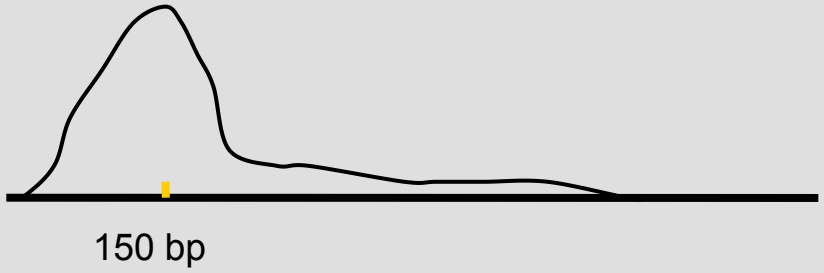
- Paprastai diskriminantinė analizė vykdoma trimis žingsniais
  - Diskriminantinio modelio pasirinkimas
  - Parametrų “apmokymas”
  - Apmokyto diskriminatoriaus naudojimas
- Parametrų “apmokymas”: geriausius klasifikavimo rezultatus duodančių parametrų įvertinimas

Neuroniniai tinklai taip pat gali būti labai tinkami atskirti vieną klasę nuo kitos



Egzonų radimas

- Klasifikatorius gali būti apmokytas atskirti egzonus nuo neegzonų panaudojant daugiau duomenų
- Genų radimo programos paprastai naudoja daugiau parametrų:



Egzonų ilgio pasiskirstymas

Egzonų radimas

Kiekvienas stačiakampis vaizduoja egzoną

Paprastai vienam realiam egzonui surandami keli persidengiantys egzonai

Todėl norint pasirinkti teisingą egzoną reikia panaudoti aukštesnio lygio informaciją, pvz. skaitymo rėmelių suderinamumą tarp gretimų egzonų

Genų radimas naujuose genomuose

- Dikodonų (heksamerų) dažnumai yra priklausomi nuo genomo – todėl vienam genomui skirtas įrankis negali būti tiesiogiai panaudotas kitam genomui

Name	ala	arg	asn	asp	cys	glu	gln	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	9.5	4.1	4.3	5.3	1.2	6	4.8	6.5	2	6.5	11.5	6	2.6	3.7	3.5	6.2	5	1.1	2.7	6.5
arg	7.9	5.5	3.9	5.3	1.1	6	5.5	5.9	2.6	6.5	11.4	5	2.2	4.7	3.6	5.5	4.4	1.4	4	6.6
asn	9.6	4.9	4.2	4.9	1	5.3	5.6	7.4	2.3	6	10	4.9	2	3.5	5.1	6.1	5.5	1.5	3.1	6.1
asp	9.3	4	4.7	5.1	1	6.7	2.9	7	1.8	7.1	9.6	6.3	2.3	4.3	3.9	5.9	5.1	1.6	3.6	6.6
cys	8.4	4.8	3.3	5.4	1.7	5.6	5.2	8.1	4.3	5.4	10.2	3.8	1.8	4.1	4.5	6.3	4.3	1.6	3.4	6.8
glu	9.4	5.8	3.6	4.5	0.8	4.9	7	5.8	2.6	5.9	12.7	5	2.4	4	3.5	5.4	5	1.1	2.8	6.8
gln	10.3	4.9	3	4.4	0.9	4.5	6.8	7	2.7	5.5	12.8	4.1	2	3.9	3.8	5.8	5.3	1.4	3	6.9
gly	8.1	4.8	3.9	5.1	1.2	6	4.6	6.4	2.4	6.8	10.5	5.8	2.7	4.8	2.4	5.8	5.1	1.4	3.7	7.5
his	7.3	4.7	4	4.8	1.5	4.9	5.6	6.9	3	6.2	10.8	4.8	1.6	5	5.2	6.8	4.9	1.7	4.2	5.1
ile	11	4.7	4.9	6.5	1.1	6.9	3.6	7.2	2.1	5.3	8.6	5.3	1.8	3.2	4.2	7	5.6	0.9	2.9	6.1
leu	10.4	4.2	4.3	5.2	1.1	5.2	3.7	6.8	2	5.6	10.6	5.3	2.3	3.8	4.5	7.4	6.2	1	2.6	6.6
lys	10.6	5.2	3.8	5.2	0.5	5.3	5.9	6.6	2.6	5.2	11.3	4.7	1.9	2.8	4.6	6	5.5	1.2	2.6	7.6
met	10.8	4.8	3.8	4.6	0.7	4.6	4.9	7	1.7	4.7	11.4	5.2	2.8	3.3	5.1	7.4	6.3	0.9	2	6.8
phe	9.6	3.7	5.2	6.5	1.2	6.4	2.7	7.9	1.9	6.7	7.4	5	2.5	3.9	3.6	8	5.8	1.3	3.3	6.3
pro	8.4	3.6	4.6	5.4	0.7	7.6	5.2	5.4	2.3	6.1	11.2	5.5	2.4	4.2	2.8	6.5	5.4	1.4	2.9	7.5
ser	9.1	4.6	3.7	5	1	5.4	5.2	7.2	2.6	6	11.6	4.5	2.2	4.1	4.1	6.5	5	1.2	3.2	6.8
thr	9.1	4.2	3.7	5.6	0.9	5.7	5.7	7.5	2.2	5.5	12	4.2	2	3.5	5.5	6.2	5.3	1.1	2.6	6.7
trp	7.1	6.3	3.2	4.8	1.3	3.9	8.5	6.6	3.6	5	14.2	3.2	2.4	4.6	3.9	5.8	4.3	1.3	3	6.1
tyr	7.9	6.5	3.6	4.9	1.2	4.5	7	7.1	2.6	5	11.7	4	1.6	4.7	4.9	6.4	4.6	1.5	3.4	5.7
val	9.6	4.1	4.4	5.9	1	6.2	3.4	6.4	1.8	6.5	10.2	5.2	2.5	3.7	3.8	7.2	6.1	1.1	2.7	7.1

shewanella

jautis

Name	ala	arg	asn	asp	cys	glu	gln	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala	11.4	5.9	3.1	4.5	1.9	5.8	3.6	7.7	1.9	4.3	9.7	4.3	2.1	3.7	6.4	6.4	5.6	1.1	2.6	6.8
arg	8.5	7.7	4	4.6	2.3	5.9	3.8	7.6	2.5	4.4	9.2	5	1.7	4	5.3	6.3	5	1.5	3.4	6.5
asn	6.3	4.9	4.9	4.4	2.1	5.3	4.1	6.9	2.2	5.6	9.7	5.4	2.1	4.1	5.9	7.3	5.3	1.9	4.6	6.2
asp	7.4	4.9	3.5	5.4	2.4	6.6	3.4	7.4	2.1	5.4	9.5	4.7	2	4.4	5.4	6.8	5.7	1.6	4	6.4
cys	6.9	5.9	4	5.4	2.7	5.6	4.9	7.1	3	4.4	8.8	5.4	1.6	3.5	6.8	7.4	5.7	1.4	2.7	5.7
glu	7.8	5.3	4.3	6.4	1.9	9.7	3.7	6.8	2	5.1	8.2	6.2	2.2	3.3	4.8	5.3	5.4	1.2	3.2	6.2
gln	7.9	5.6	4.2	5	2	6.6	5.1	6.9	2.1	4.7	9.3	5.7	2	3.3	5.9	5.7	6.1	1.6	3.3	6.2
gly	7.9	5.8	3.9	5	1.9	6.2	3.5	8	1.8	4.7	8.7	5.2	1.7	3.7	6.9	7.4	5.8	1.4	3.2	6.2
his	6	5.8	4.3	3.5	2.9	5.1	4.1	6.3	3.2	4.5	10.6	4.8	1.6	4.5	6.7	6.6	6.1	1.7	3.9	6.9
ile	6.2	4.9	4.9	4.7	2.4	5.3	4.6	5.8	2.2	6	9.9	5.3	2.1	4.1	5.3	7.7	6.9	1.2	3.7	6
leu	7.7	5.6	4.1	4.7	2.1	5.8	4.5	6.8	2.1	4.6	11	5.4	1.9	3.7	5.7	7	5.5	1.2	3.1	6.4
lys	6.3	5.2	4.8	5.2	2.1	7.2	3.7	6.7	2.2	6	8.5	7.5	2	3.5	4.8	6.1	5.8	1.6	3.5	6.3
met	9.3	5.3	4.1	5.9	1.6	6.1	3.5	6.4	1.6	4.1	9.6	6.6	2.6	4	5.1	6.9	5.5	1	3.2	6.6
phe	6	5.4	4.5	5.2	2.5	5.5	4.1	6.5	2.3	5.3	10.2	5.2	1.8	4.1	5.3	7.8	5.8	1.4	3.9	6.2
pro	8.5	5.4	3.1	5.1	1.9	6.7	3.9	9.5	1.9	4.3	7.7	4.3	1.7	3.3	8.7	6.9	5.7	1.4	2.8	6.4
ser	6.7	5.4	3.8	4.9	2.3	5.4	4	7.9	2.1	4.5	9.5	5.2	1.8	4	5.7	8.6	6.2	1.4	3	6.4
thr	7.5	4.6	3.7	5	2.6	5.7	3.8	6.8	2	5.2	9.7	4.4	1.8	3.9	6	7.2	7.3	1.5	3.5	6.9
trp	7.1	5.2	4.9	5.5	2.3	5.4	4.3	5.8	2.2	5.6	9.5	6.6	2.1	3.8	4.1	6.4	5.9	1.7	3.7	6.8
tyr	5.8	5.7	5	5.1	2.3	5.7	4.1	6.2	2.4	5	8.6	5.6	1.9	5	4.8	6.7	6.3	1.5	4.8	6.5
val	7.6	5	4.4	5.2	2.4	5.7	3.7	6.3	1.9	5	9.3	5.1	2.1	4.1	5.5	6.9	6.6	1.1	3.6	7.4

Genų radimas naujuose genomuose

- Prieš apmokant genų paieškos variklį buvo padaryta prielaida, kad yra žinomi kai kurie genai ir “ne-genai”

- O tai ką daryti, jei reikia surasti genus ką tik nusekvenuotam genomui?

Genų radimas naujuose genomuose

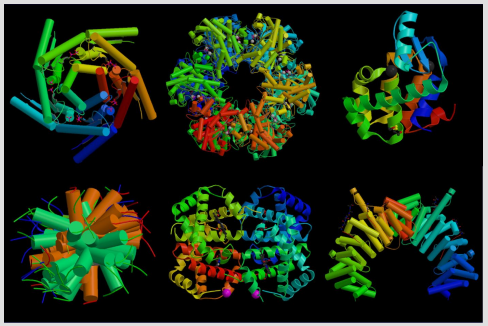
- Dalį genų galime surasti naudodami homologijų paiešką, ieškant žinomų genų iš GenBank'o
  - BLAST, FASTA ir kt.
- Taigi randame koduojančius regionus
- Nekoduojančius regionus surandame sakydami kad tarpgeniniai regionai yra nekoduojantys
- Ir tada galime apmokyti paieškos variklį

Populiariausios genų paieškos programos

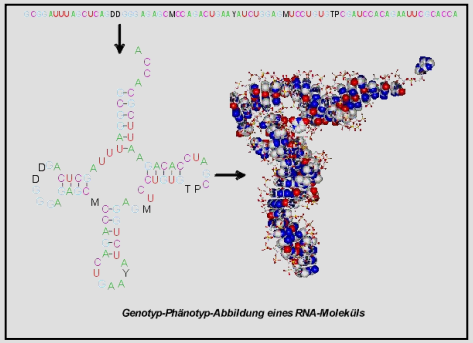
- GeneScan
  - Žmogaus genomas
- GRAIL
  - Žmogaus ir kitų eukariotų genomai
- GeneMark
  - Bakteriniai genomai
- Glimmer
  - Bakteriniai genomai

Problemos, kurias dar reikia išspręsti

- RNR genų radimas
  - Dažniausiai remiasi RNR antrinėmis struktūromis



baltymas



RNA

- Alternatyvaus “spliceing”’o spėjimas