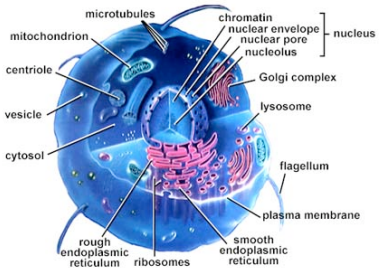
Sequence Analysis

BBSI 2006: Lecture #(χ+1)

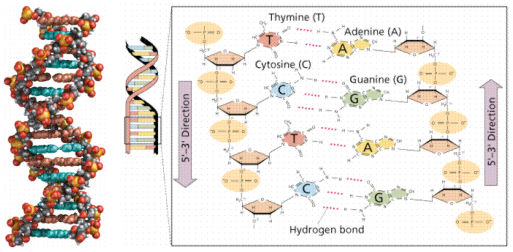
Takis Benos (2006)

Molecular Genetics 101

Cell's internal world



DNA - Chromosomes - Genes



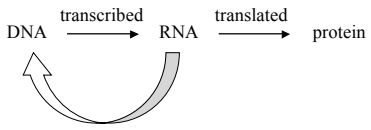
What is a "gene"?

- We cannot define it (but we know it when we see it…)
- A loose definition:

"Gene" is a *DNA/RNA* *information unit* that is able to perform a function in a cellular environment

Protein coding genes

Central Dogma:



Open Reading Frames (ORFs)

```
aatagcgaat tttccaacga caaaagctaa atatcgcaaa aacctcagta aaaatcttgc  60
tggagctatt attgctaagt aacatttacc ccctgaagtt aatggatcaa tcaagagaga 120
tgtgggctgt aATGaatcgt cttattgaat taacaggttg gatcgttctt gtcgtttcag 180
tcattcttct tggcgtggcg agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagTAAgct ctgcacttgt ggagcgacat gctgcccgtc cgggtgcatg 300
ttttcacttg tcggatatta aaccaggaat ttattatctt gttcgatgtt gtaataaa   358
```

Open Reading Frames (ORFs)

```
aatagcgaat tttccaacga caaaagctaa atatcgcaaa aacctcagta aaaatcttgc  60
tggagctatt attgctaagt aacatttacc ccctgaagtt aatggatcaa tcaagagaga 120
tgtgggctgt aATGaatcgt cttattgaat taacaggttg gatcgttctt gtcgtttcag 180
tcattcttct tggcgtggcg agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagTAAgct ctgcacttgt ggagcgacat gctgcccgtc cgggtgcatg 300
ttttcacttg tcggatatta aaccaggaat ttattatctt gttcgatgtt gtaataaa   358
```

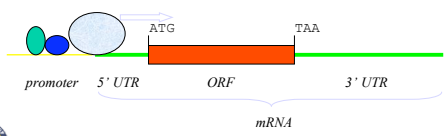
M N R L

K stop

```
MNRLIELTGWIVLVVSVILLGVASHIDNYQPPEQSASVQHK
```

Gene's characteristics

```
aatagcgaat tttccaacga caaaagctaa atatcgcaaa aacctcagta aaaatcttgc  60
tggagctatt attgctaagt aacatttacc ccctgaagtt aatggatcaa tcaagagaga 120
tgtgggctgt aATGaatcgt cttattgaat taacaggttg gatcgttctt gtcgtttcag 180
tcattcttct tggcgtggcg agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagTAAgct ctgcacttgt ggagcgacat gctgcccgtc cgggtgcatg 300
ttttcacttg tcggatatta aaccaggaat ttattatctt gttcgatgtt gtaataaa   358
```

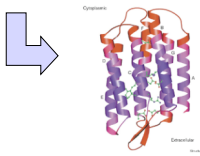


Transcription

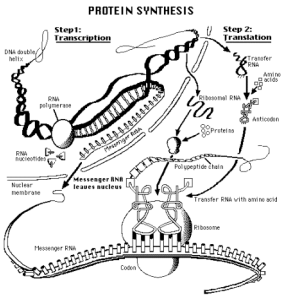
```
                                        cugaaguu aauggaucaa ucaagagaga 120
uguggGcugu aAUGaaucgu cuuauugaau uaacagguug gaucguucuu gucguuucag 180
ucauucuucu uggcguggcg agucacauug acaacuauca gccaccugaa cagagugcuu 240
cgguacaaca caagUAAgcu cugcacuugu ggagcgacau gcugcccguc cgggugcaug 300
uuuucacuug ucggauauua aaccaggaau uuauuaucuu guucgauguu guaauaaa   358
```

Translation


```
MNRLIELTGWIVLVVSVILLGVASHIDNYQPPEQSASVQHK
```

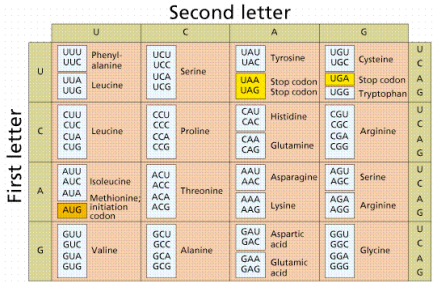


Protein coding genes (cntd)



Source: http://www.ems.maricopa.edu/faculty/farabee/BIOBK/BioBookPR OTSYn.html

Genes and Proteins

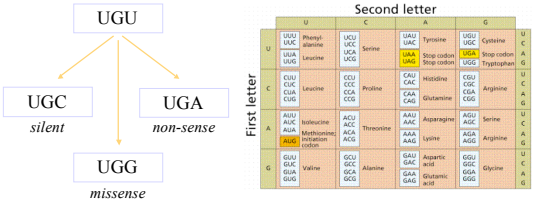


Alterations of the DNA

Base substitutions:

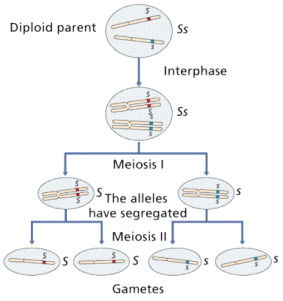
- silent: no a.a. replacement
- missense: a.a. replacement
- non-sense: a.a. → stop codon replacement

Alterations of the DNA (cntd)



Source: http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookPROTSYn.html

Inheritance - genetic differences



Molecular evolution

Two species will acquire mutations proportionally to their divergence time. However:

- all proteins do not change in the same pace
- a given protein does not necessarily change in the same pace throughout time
- different parts of the same protein change at different paces

Molecular evolution (cntd)

Human (C11A_HUMAN; P05108) *vs. Drosophila* (C11A_PIG; P10612)

```
Query: 1   MLAKGLPPRSVLVKGYQTFLSAPREGLGRLRVPTGEGAGISTRSPRPFNEIPSPGDNGWL 60
           ML+tGL  RSVLVKG Q FLSAPRE  G  RV TGEGA iSTrtPRPFtEIPSPGDNGWt
Sbjct: 1   MLARGLALRSVLVKGCQPFLSAPRECpGHPRVGTGEGACISTKTPRPFSEIPSPGDNGWI 60

Query: 61  NLYHFWRETGTHKVHLHHVQNFQKYGPIYREKLGNVESVYVIDPEDVALLFKSEGPNPER 120
           NLY FW+E GT K+H HHVQNFQKYGPIYREKLGN+ESVY+IDPEDVALLFK EGPNPER
Sbjct: 61  NLYRFWKEKGTQKIHYHHVQNFQKYGPIYREKLGNLESVYIIDPEDVALLFKFEGPNPER 120

Query: 121 FLIPPWVAYHQYYQRPIGVLLKKSAAWKKDRVALNQEVMAPEATKNFLPLLDAVSRDFVS 180
           + IPPWVAYHQ+YQRPVGVLLKKS AWKKDR+ LN EVMAPEA KNF+PLLD VS+DFV
Sbjct: 121 YNIPPWVAYHQHYQRPVGVLLKKSGAWKKDRLVLNTEVMAPEAIKNFIPLLDTVSQDFVG 180

Query: 181 VLHRRIKKAGSGNYSGDISDDLFRFAFESITNVIFGERQGMLEEVVNPEAQRFIDAIYQM 240
           VLHRRIK+ GSG +SGDI +DLFRFAFESITNVIFGER GMLEE+V+PEAQ+FIDA+YQM
Sbjct: 181 VLHRRIKQQGSGKFSGDIREDLFRFAFESITNVIFGERLGMLEEIVDPEAQKFIDAVYQM 240

Query: 241 FHTSVPMLNLPPDLFRLFRTKTWKDHVAAWDVIFSKADIYTQNFYWELRQKGSVHHDYRG 300
           FHTSVPMLNLPPDLFRLFRTKTW+DHVAAWD IF+KA+ YTQNFYW+LR+K     ++Y G
Sbjct: 241 FHTSVPMLNLPPDLFRLFRTKTWRDHVAAWDTIFNKAEKYTQNFYWDLRRKRE-FNNYPG 299

Query: 301 MLYRLLGDSKMSFEDIKANVTEMLAGGVDTTSMTLQWHLYEMARNLKVQDMLRAEVLAAR 360
           +LYRLLG+ K+  ED+KANVTEMLAGGVDTTSMTLQWHLYEMAR+L VQ+MLR EVL AR
Sbjct: 300 ILYRLLGNDKLLSEDVKANVTEMLAGGVDTTSMTLQWHLYEMARSLNVQEMLREEVLNAR 359

Query: 361 HQAQGDMATMLQLVPLLKASIKETLRLHPISVTLQRYLVNDLVLRDYMIPAKTLVQVAIY 420
           QAQGD+ MLQLVPLLKASIKETLRLHPISVTLQRYLVNDLVLRDYMIPAKTLVQVAIY
Sbjct: 360 RQAQGDTSKMLQLVPLLKASIKETLRLHPISVTLQRYLVNDLVLRDYMIPAKTLVQVAIY 419

Query: 421 ALGREPTFFPDPENFDPTRWLSKDKNITYFRNLGFGWGVRQCLGRRIAELEMTIFLINML 480
           A+GR+P FF +P  FDPTRWL K++++ +FRNLGFGWGVRQC+GRRIAELEMT+FLI++L
Sbjct: 420 AMGRDPAFFSNPGQFDPTRWLGKERDLIHFRNLGFGWGVRQCVGRRIAELEMTLFLIHIL 479

Query: 481 ENFKVEIQHLSDVGTTFNLILMPEKPISFTFWPFNQEATQ 520
           ENF+VE+QH SDV T FNLILMP+KPI  F PFNQ+  Q
Sbjct: 480 ENFKVELQHFSDVDTIFNLILMPDKPIFLVFRPFNQDPLQ 519
```

Molecular evolution (cntd)

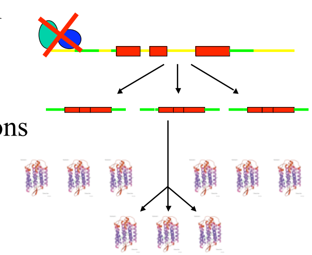
Human (C11A_HUMAN; P05108) *vs.* zebrafish (Cyp11a1; Q8JH93)

```
Query: 34  TGEGAGISTRSPRPFNEIPSPGDNGWLNLYHFWRETGTHKVHLHHVQNFQKYGPIYREKL 93
           T  G      + +PFN+IP    N  L++  F +  G   VH   V NF+ +GPIYREK+
Sbjct: 27  TRSGRAPQNSTVQPFNKIPGRWRNSLLSVLAFTKMGGLRNVHRIMVHNFKTFGPIYREKV 86

Query: 94  GNVESVYVIDPEDVALLFKSEGPNPERFLIPPWVAYHQYYQRPIGVLLKKSAAWKKDRVA 153
           G  +SVY+I PED A+LFK+EG +P R  +  W AY  Y  +  GVLLK+  AWK DR+
Sbjct: 87  GIYDSVYIIKPEDGAILFKAEGHHPNRINVDAWTAYRDYRNQKYGVLLKEGKAWKTDRMI 146

Query: 154 LNQEVMAPEATKNFLPLLDAVSRDFVSVLHRRIKKAGSGNYSGDISDDLFRFAFESITNV 213
           LN+E++ P+    F+PLLD V +DFV+ ++++I+++G   ++ D++ DLFRF+ ES++ V
Sbjct: 147 LNKELLLPKLQGTFVPLLDEVGQDFVARVNKQIERSGQKQWTTDLTHDLFRFSLESVSAV 206

Query: 214 IFGERQGMLEEVVNPEAQRFIDAIYQMFHTSVPMLNLPPDLFRLFRTKTWKDHVAAWDVI 273
           ++GER G+L + ++PE Q FID +  MF T+ PML LPP L R   +  WK+HV AWD I
Sbjct: 207 LYGERLGLLLDNIDPEFQHFIDCVSVMFKTTSPMLYLPPGLLRSIGSNIWKNHVEAWDGI 266

Query: 274 FSKADIYTQNFYWELRQKGSVHHDYRGMLYRLLGDSKMSFEDIKANVTEMLAGGVDTTSM 333
           F++AD   QN + + ++     +  Y G+L  LL   K+S EDIKA+VTE++AGGVD+ +
Sbjct: 267 FNQADRCIQNIFKQWKENPEGNGKYPGVLAILLMQDKLSIEDIKASVTELMAGGVDSVTF 326

Query: 334 TLQWHLYEMARNLKVQDMLRAEVLAARHQAQGDMATMLQLVPLLKASIKETLRLHPISVT 393
           TL W LYE+AR   +QD LRAE+ AAR   +GDM  M++++PLLKA++KETLRLHP++++
Sbjct: 327 TLLWTLYELARQPDLQDELRAEISAARIAFKGDMVQMVKMIPLLKAALKETLRLHPVAMS 386

Query: 394 LQRYLVNDLVLRDYMIPAKTLVQVAIYALGREPTFFFDPENFDPTRWLSKDKNITYFRNL 453
           L RY+  D V+++Y IPA TLVQ+ +YA+GR+  FF  PE + P+RW+S ++   YF++L
Sbjct: 387 LPRYITEDTVIQNYHIPAGTLVQLGVYAMGRDHQFFPKPEQYCPSRWISSNRQ--YFKSL 444

Query: 454 GFGWGVRQCLGRRIAELEMTIFLINMLENFRVEIQHLSDVGTTFNLILMPEKPISFTFWP 513
           GFG+G RQCLGRRIAE EM IFLI+MLENFR+E Q   +V + F L+LMPEKPI  T  P
Sbjct: 445 GFGFGPRQCLGRRIAETEMQIFLIHMLENFRIEKQKQIEVRSKFELLLMPEKPIILTIKP 504

Query: 514 FN 515
            N
Sbjct: 505 LN 506
```

BBSI 2006 26-MAY-2006

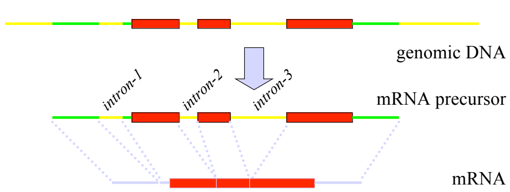
Transcription regulation

- promoter region
- expression rates
- degradation
- post modifications



BBSI 2006 26-MAY-2006

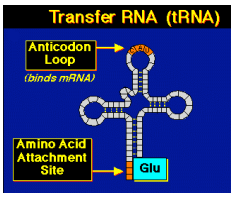
Splicing



BBSI 2006 26-MAY-2006

Non-coding genes

- tRNA
- ribosomal RNA
- snoRNA
- microRNA
- etc



Source: http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookPROTSYn.html

Elements of Probability Theory (with examples)

Outline

- Conditional Probabilities
- Markov Chains
- Hidden Markov Models
- Information measures

Probabilities

Definition:

$$P(x) = \frac{\text{\# favourable outcomes } (x)}{\text{total \# possible outcomes}}$$

Conditional Probabilities

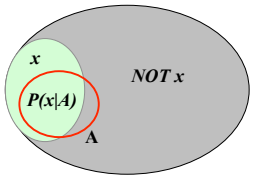
Definition:

$$P(x|A) = \frac{\text{\# favourable outcomes (x) given A}}{\text{total \# possible outcomes given A}}$$

Conditional Probabilities (cntd)

Example:

- *outcome x*: *tomorrow I'll go hiking*
- *event A*: *tomorrow will be sunny*



Conditional Probabilities (cntd)

- *Joint probability*: $P(X,Y) = P(X|Y)\,P(Y)$
- If $P(X|Y) = P(X)$ then X,Y *independent*

$$P(X,Y) = P(X)\,P(Y)$$

- *Marginal probability:*

$$P(X) = \sum_{Y} P(X,Y) = \sum_{Y} P(X \mid Y)P(Y)$$

Conditional Probabilities (cntd)

- *Bayes' theorem*

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} = \frac{P(Y \mid X)P(X)}{\sum_{X} P(Y \mid X)P(X)}$$

- *Posterior probabilities are the compromise between data and prior information.*

Bayes: Application-1

- Problem (from Durbin *et al.*, *1998*):

A rare genetic disease is discovered with population frequency one in 1 million. An extremely good genetic test is 100% sensitive (always correct if you have the disease) and 99.99% specific (false positive rate 0.01%). Will you be willing to take such a test?

- *Hint:* What is the probability that you have the disease, if the test is positive?

Bayes: Application-1 (cntd)

- Answer:

$P(D \mid +) = P(+ \mid D)\, P(D) / P(+) =$

$= 1.0 * 10^{-6} / [1.0 * 10^{-6} + 10^{-4} * (1 - 10^{-6})] =$

$= 0.0099$

Application of Bayes-2

- Problem:

Given a set of transmembrane proteins with specified membrane domains of length *L (training set)*, can you develop a probabilistic model that predicts which parts of a new transmembrane protein are likely to be membrane domains?

Application of Bayes-2 (cntd)

- Solution:
  - Suppose that we suspect that the amino acid frequencies differ between membrane and non-membrane regions.
  - Using the *training* set, calculate the probabilities, $P(a_i|D)$, that each amino acid a_i is part of a membrane domain (D). Also, using the non-membrane parts, calculate the corresponding probabilities, $P(a_i|\text{not D})$.

Application of Bayes-2 (cntd)

- Solution *(cntd)*:
  - Divide the new protein into segments.
  - Using Bayes theorem, calculate the posterior probability of each segment being a membrane domain using the P(a_i|D).

$$P(X \mid M); M := \arg\max_M \frac{P(M \mid D)P(D)}{\sum_d P(M \mid d)P(d)}$$

Markov chains

- What is a Markov chain?

  - Markov chain of order *n* is a stochastic process of a series of outcomes, in which the probability of outcome *x* depends on the state of the previous *n* outcomes.

Markov chains (cntd)

- Markov chain (of *first* order):

$$P(x) = P(X_L, X_{L-1}, \ldots, X_1) =$$
$$= P(X_L \mid X_{L-1}, \ldots, X_1)P(X_{L-1} \mid X_{L-2}, \ldots, X_1)\ldots P(X_1) =$$
$$= P(X_L \mid X_{L-1})P(X_{L-1} \mid X_{L-2})\ldots P(X_2 \mid X_1)P(X_1) =$$
$$= P(X_1)\prod_{i=2}^{L} P(X_i \mid X_{i-1})$$

- *Transition probabilities*: $P(X_i \mid X_{i-1})$

Application of Markov chains

- Problem (from Durbin *et al.*): CpG islands

Given two sets of sequences from the human genome, one with CpG islands and one without, can you calculate a model that can predict the CpG islands?

Application of Markov chains (cntd)

- Solution:

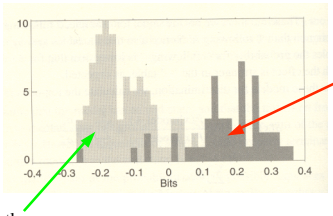
+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

Application of Markov chains (cntd)

- Histogram of scores (CpG islands):



Hidden Markov Models

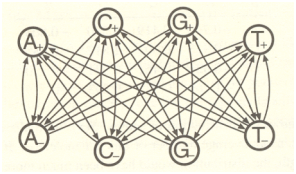
- What is a Hidden Markov Model?
  - A Markov process in which the probability of an outcome depends also in a (hidden) random variable (state).
  - *Transition* probability: the probability of reaching a state given the previous state.
  - E*mission* probability: the probability of an outcome given the state.

Hidden Markov Models (cntd)

- Graphical representation of the HMM:

CpG islands

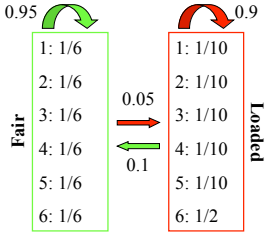
(transition probabilities)



- *Question:* Where is the Markov process here?

Application of HMMs

- Problem (from Durbin *et al.*): dishonest casino



Application of HMMs (cntd)

- Problem (from Durbin *et al.*): dishonest casino

> Given (1) the previous model and (2) a series of die rolls (x_i, i=1,…,L), can we predict which of the rolls are coming from the fair and which from the loaded die?

- *<u>Question</u>:* What is "hidden" here?

Application of HMMs (cntd)

- Answer: YES
  - Viterbi algorithm (best path)
  - Forward-backward algorithm (probability of state k in outcome x_i)

HMMs: Viterbi algorithm

- Viterbi predictions: 300 rolls of die

```
Rolls   315116246446644245311321631164152133625144543631656626566666
Die     FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLL
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLL

Rolls   651166453132651245636664631636663162326455236266666625151631
Die     LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLFFFFFFFFF
Viterbi LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF

Rolls   222555441666566563564324364131513465146353411126414626253356
Die     FFFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLL
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL

Rolls   366163666466232534413661661163252562462255265252266435353336
Die     LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Rolls   233121625364414432335163243633665562466662632666612355245242
Die     FFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFFFFF
Viterbi FFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLFFFFFFFFFFFF
```

HMMs in biology

- General comments:

- Usually the structure of the model is unknown
- The transition and emission probabilities are calculated based on trusted training set(s) and the postulated model
- Predictions are based on the Viterbi or the forward-backward algorithm, depending on the question asked

Information measures

- Definitions:
  - Entropy: $\mathrm{H}(P) = \mathbf{E}(-\log P) = -\sum_{i=1}^{n} p_i \log p_i$
  - Relative Entropy: $\mathrm{H}(P,Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$
  - Mutual Information: $M(X,Y) = \sum_{i,j} P(x_i,y_j) \log \frac{P(x_i,y_j)}{P(x_i)P(y_j)}$