Sequence Analysis

BBSI 2006: Lecture #(χ+1)

Takis Benos (2006)
Molecular Genetics 101
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:

DNA \xrightarrow{\text{transcribed}} RNA \xrightarrow{\text{translated}} \text{protein}

\textcircled{DNA} \xrightarrow{\text{transcribed}} \RNA \xrightarrow{\text{translated}} \text{protein}
Open Reading Frames (ORFs)

aatagcgaat tttcaacga caaaagctaa atatcgcaaa aacctcagta aaaatc tgtgc
60
tgagctatt attgctaatg aacattttacc ccctgaagtt aatggatcaa tcaagagaga 120
tggtcgctgt aATGaatcgt cttattgaat taacaggttg gatcgttctt gtcgtttcag 180
tcattttct tggcgtgccc agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtacaaca caagTAAgct cttggcgttgt ggagcgacat gctgcccgtc cgggtgcatg 300
cttctttct tggcgtgccc agtcacattg acaactatca gccacctgaa cagagtgctt 358
Open Reading Frames (ORFs)

aatagcgaat tttccaacga caaaagctaa atatcgcaaa aacctcagta aaaaatcttg 60
tggagctatt attgctaatgt aacattttacc ccctgaagtt aatggatcaaa tcaagagaga 120
tgtgggctgt aATGaatcgt cttattgaat taacaggttg gatcgttctt gtcgttttcag 180
tcatttttcct tggcgtggcgc agtcacattg acaactatca gccacctgaa cagagtgctt 240
cggtaacaaca caagTAAgct ctgcactttgat ggagcgacat gctgcccgcgc cgggtgcgtg 300
ttttcactttg tcggatatta aaccaggaat ttattatctt gttcgatgtt gtaataaaa 358

MNRLIELTGWIVLVSVIILGVASHIDNYQPPEQSAVQHK

BBSI 2006 26-MAY-2006

© 2006 P. Benos
Gene’s characteristics

aataggcgaat tttccaaacga caaaaagctaa atatcgcaaa aacctcagta aaaaatcttgc 60
tggagctatt attgctaagt aacattttacc ccctgaagtt aatggatcaa tcaagagaga 120
tgtgggctgt aATGaatcgt cttattgaat taacaggttg gatcgttcctt gtcgtttcag 180
tcattctctct tggcggtggcg agtcacatgg acaactatca gccacctgaa cagagtgcctt 240
cggtacaaca caagTAAgct ctgcacctgtt gtagcgcacat gctgcgccgctc cggtgctgcatg 300
ttttcactttg tcggatatatt aaccaggaat ttatttatcttt gttcgatgttt gtaataaa 358

promoter 5’ UTR ORF 3’ UTR mRNA
Transcription

cugaaguu aauggaucaa ucaagagaga 120
ugugggcugu aAUGaaucgu cuuauugaaau uaacagguug gaucguucuu gucguuuucag 180
ucauuucuucu uggcguggcg agucacauug acaacuaauca gccaccugaa cagagugcuu 240
cgguacaaca caagUAAgcu cugcakuugu gggagcgacau gcugcccguc cgggugcaug 300
uuuuacuug uccgauuaau aaccaggaau uuauuaucuu guuucgauguu guaauaaa 358

CAP...aaauu... Poly(A) tail
Translation

cuga...

...aaauaaa —aaaaaaa

MNRLIELTGWIVLVVSIVLLGVASHIDNYQPPEQSAVSQHK
Protein coding genes (cntd)

PROTEIN SYNTHESIS

Step 1: Transcription
- DNA double helix
- RNA polymerase
- RNA nucleotides

Messenger RNA leaves nucleus
- Nuclear membrane

Step 2: Translation
- Transfer RNA
- Amino acids
- Ribosomal RNA
- Anticodon

Ribosomes
- Polypeptide chain
- Transfer RNA with amino acid

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

BBSI 2006 26-MAY-2006

© 2006 P. Benos
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)

- **UGU**
  - **UGC** (silent)
  - **UGA** (non-sense)
  - **UGG** (missense)

Source: [http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookPROTSYN.html](http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookPROTSYN.html)
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. Pig (C11A_PIG; P10612)

Query: 1  MLAKGLPPRSVVLKGYQTFLSAPREGLRLRVPTEGAGAGISTRSPRPNEIPSPGDNGWL
Sbjct: 1  MLARGLALRSVVLKGCQPFLSAPRECPGHPRVTGEGACISTKTPRPSEIFPSGDNGW

Query: 61  NLYHWREGTGTHKHLHVQNFQKYPIYREKLGVESVYVIDPEDVALLFKSEGPNPER
Sbjct: 61  NLYRFWKEKGTQIKHYHVQNFQKYPIYREKLGNLESVYIIPLEDVALLFKFEGPNPER

Query: 121  FLIPPWVAYHQYQRPIGVVLLKSSAWKDRVNLNQVEVMAPEATKKNFLPLLQAVSRDFS
Sbjct: 121  YNIPPVAYHQKPKGVVLKSSAWKKDRLVNLTEVMAPEAIKKNFIPPLTVSDFVG

Query: 181  VLHRRIKKAGSGNSGDISSDLRFARAFESITNVIKERQGMLEEVVNPAREQFIDAVIYQM
Sbjct: 181  VLHRRIKQQSGKFSDIREDLRFARAFESITNVIKERQGMLEEVVNPAREQFIDAVIYQM

Query: 241  FHTSVPLNLPDLLFRLFRKTWKDHHAVAWDFVSKADIYTNFSGWELRQKSGVHHDYRG
Sbjct: 241  FHTSVPLNLPDLLFRLFRKTWDRHDAWVDIFNKAETYQNFYGWDLRKRE-FNNYPG

Query: 301  MLHRDGKDSFKSEDIKANVTEMLAGGVDTTSMTLQWHLYEMARNLKQVQMLRAEVLAR
Sbjct: 300  ILYRLLGDSKLLSEVDKANVTEMLAGGVDTTSMTLQWHLYEMARSLNQMLEELVNL

Query: 361  HQAGSQDMATMLQLVPLKLASKIKETLRLHPISVTLYVNLVLRDYMPAKTLVQVAIY
Sbjct: 360  RQAQGDTSKMLQVLPLKLASKIKETLRLHPISVTLYVNLVLRDYMPAKTLVQVAVY

Query: 421  ALGREQTFDPFPENPDRTPWLSKDKNTYFRNLGFGWCVQCGLGRRIAELEMTEFLINML
Sbjct: 420  AMGRDPAFFSNPGFDTPRGLKEDLIHRNLFGWGVQCGRRIAELEMTEFLHIL

Query: 481  ENFRVEIQHLSDGVTFNLILMEPKPISTFVFPFNOQEAATQ
Sbjct: 480  ENFKVELQHFSVDETFNLILMMPDKPIFLVFRPNOQDPQL

© 2006 P. Benos
Molecular evolution (cntd)

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

Query: 34  TGEAGAGISTRSPRPFNEIPSPGDNGLNLYHFWRETGTHKVHLHHVNQNFQKYGPIYREKL  93
Sbjct:  27  TRSGRAPQNSTVQPFNKIPGRWRNSSLVLAFTKMGGLRNVHIRMVHNFLTFGPIYREKV  86
Query: 94  GNVESVYVIDPEDVALLFKSEGPNERPFLIPPVAYHQQYQRPIVGVLKKSAAWKKDRVA 153
Sbjct:  87  GIYDSAIIKPEDGAILFKAEGHHPNKNRDVTAYRDYRNRQKYGVLLLKEGKAWKTDRI  146
Query: 154  LNQVEVMAEATKNNLPLLDDAVSRDFVSVLHHRKKAGSGNGSYSGDISDLLFRAFESITNV 213
Sbjct: 147  LNKELLPLQKGTFVPLDEVGQDFVARVKNKIEQSGKQWTDLTHDLFRFSESVAV  206
Query: 214  IFGERQGMLEEVNVPAAQRFDAYMQFHTSVPMNLDPDLLFRLFRKTWADHVAAGWDV 273
Sbjct: 207  LYGERLGLDNDPEFQMDFICVMSVMFLLPPLRLGRSISNIWKHVVWADGI  266
Query: 274  FSKADIYTQNYRELRQQGVSVVHDYLRLGDKSMFEDKIANVETMAGVVTSTSM 333
Sbjct: 267  FNQADRCIQNFQKCNKENPEGNGKYPGVLAILLMQDLISIEDIKASVELLAMGVSSTV  326
Query: 334  TLQWNLVEARNLKVQDMLDRAELAHRQAQDMDATMLQLPLKASIKETLHRHPSVT 393
Sbjct: 327  TLLWTLYERALQDQLDELRQASIARIFGMQVKMIPILLKAKTELRLHPWSMV 386
Query: 394  LQRYLVLNDLVRDYMPAKTQLVQAIYALGREATFFFDPENFDPRWLSKDKNTYFRNL 453
Sbjct: 387  LRPRYTEDTVIQNYHIPAGFLTVQGLVAYMGRDHQFPKEVPCRSWSSNRQ--YFKSL  444
Query: 454  GFGWGVRQCLGRRIAELMTIFLINLNEFVRIEIQHLDSDVGTFNLILMPKIPSTFWP 513
Sbjct: 445  GFGFVGPRQCLGRRIAELMTIFILIHMENFRIEIQKQIEVRSKFLLLLMPKPIILTKP  504

514  FN  515
505  LN  506
Gene expression regulation

- promoter region
- expression rates
- degradation
- post modifications
Splicing

genomic DNA

mRNA precursor

mRNA
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:

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

Conditional probabilities:

\[ P(x|A) = \frac{\text{# favourable outcomes (}x\text{) given } A}{\text{total # possible outcomes given } A} \]
Conditional Probabilities

• 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:**

\[
P(D | +) = \frac{P(+ | D) \ P(D)}{P(+)} = \\
= 1.0 \times 10^{-6} / [1.0 \times 10^{-6} + 10^{-4} \times (1 - 10^{-6})] = \\
= 0.0099
\]
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):

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

• Transition probabilities: \( P(X_i | X_{i-1}) \)
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:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>0.171</td>
<td>0.368</td>
<td>0.274</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>0.161</td>
<td>0.339</td>
<td>0.375</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>0.079</td>
<td>0.355</td>
<td>0.384</td>
<td>0.182</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>0.300</td>
<td>0.205</td>
<td>0.285</td>
<td>0.210</td>
</tr>
<tr>
<td></td>
<td>0.322</td>
<td>0.298</td>
<td>0.078</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td>0.248</td>
<td>0.246</td>
<td>0.298</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td>0.177</td>
<td>0.239</td>
<td>0.292</td>
<td>0.292</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.740</td>
<td>0.419</td>
<td>0.580</td>
<td>-0.803</td>
</tr>
<tr>
<td></td>
<td>-0.913</td>
<td>0.302</td>
<td>1.812</td>
<td>-0.685</td>
</tr>
<tr>
<td></td>
<td>-0.624</td>
<td>0.461</td>
<td>0.331</td>
<td>-0.730</td>
</tr>
<tr>
<td></td>
<td>-1.169</td>
<td>0.573</td>
<td>0.393</td>
<td>-0.679</td>
</tr>
</tbody>
</table>
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.

  ➢ Emission 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

- Fair:
  - 1: 1/6
  - 2: 1/6
  - 3: 1/6
  - 4: 1/6
  - 5: 1/6
  - 6: 1/6

- Loaded:
  - 1: 1/10
  - 2: 1/10
  - 3: 1/10
  - 4: 1/10
  - 5: 1/10
  - 6: 1/2
Application of HMMs (cntd)

• Problem (from Durbin et al.): dishonest casino

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

• Question: 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
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:

\[ H(P) = E(-\log P) = -\sum_{i=1}^{n} p_i \log p_i \]

➢ Relative Entropy:

\[ H(P, Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \]

➢ Mutual Information:

\[ M(X,Y) = \sum_{i,j} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \]