PONENCIA # 4

TOOLBOX DE BIOINFORMATICA: ENTORNO DE SOFTWARE INTEGRADO PARA EL GENOMA Y ANÁLISIS PROTEÓMICO

JAIRO PERTUZ CAMPO
Físico, Instructor y asesor de MATLAB, E.MAILS: jpertuz@udem.edu.co, pertuzjairo@yahoo.es

SOLON PINO G.
D. SOFTWARE PARA INGENIERIA
COMPONENTES ELECTRONICAS LTDA

1. PRESENTACIÓN DEL BIOINFORMATICS TOOLBOX

Componentes Electrónicas Ltda Distribuidor Autorizado en Colombia de **The MathWorks Inc.** anunció el 6 de Noviembre la disponibilidad de su **nuevo Bioinformatics Toolbox** para MATLAB ®. El producto facilita el análisis de datos complejos de bioinformática, de las sucesiones de análisis genomático de datos en micro-arreglos, y velocidades de desarrollo y despliegue de algoritmos. Como el Toolbox se construye en MATLAB ®, los especialistas en bioinformática pueden confiar en un análisis de datos probado y un ambiente de desarrollo de algoritmos que les permiten responder eficiente y rápidamente, con menos tareas de codificación.

"MATLAB ® proporciona una abundancia de herramientas que son críticas al planear el desarrollo de un algoritmo. El Bioinformatics Toolbox se suma naturalmente a MATLAB ® manteniendo una interface sin limitaciones y accediendo con él a los almacenes de datos del genoma esenciales y de dominio público, y proporcionando los constructores de bloques necesarios para procesar los datos "genomáticos" dijo Bahram Ghaffarzadeh Kermani, Ph. D., del personal científico de Illumina, Inc. en San Diego.

El Bioinformatics Toolbox se diseña para ayudar a los usuarios a entender y visualiza cantidades grandes de datos de la investigación asociada con las aplicaciones de la tecnologia biológica o bio-tecnologia tradicionales, en un fragmento del tiempo. Con este último Toolbox The MathWorks está entregando el poder y versatilidad de su ambiente informático técnico integrado directamente a la bio-tecnologia y las industrias farmacéuticas. Aprovechando la funcionalidad y riqueza de la programación en MATLAB ®, el Bioinformatics Toolbox proporciona un ambiente abierto, extensible a científicos y a bioinformáticos para el desarrollo de algoritmos, manipulación de este tipo de datos y habilita la comunicación fácil de aplicaciones por las diferentes secciones de la investigación. Como resultado, los bioinformaticos pueden usar el toolbox para enfocar los esfuerzos en el centro de su trabajo - la investigación y análisis - sin los riesgos asociados con usar programas o software dispares.
:

Entre sus numerosos rasgos y capacidades, el Bioinformatics Toolbox proporciona el acceso a archivos del genoma en formatos normales, los bancos de datos basados en la web como GenBank y PIR, y las fuentes de los datos en línea. El toolbox también ofrece las rutinas especializadas para visualizar los datos de microarray (micro-arreglos), incluso las cajas de gráfico, los gráficos I-R y los mapas espaciales de calor.

Los rasgos importantes adicionales del Bioinformatics Toolbox incluyen:

- Archivos y expresiones en formato comprensibles en genética, genomática y proteomática.
- Conversión, adaptación y estadísticas de secuencias del genoma y proteínas.
- Funciones de análisis proteínico.
- Gráficas de puntos, gráficas de grupos, gráficas de sectores y otras representaciones gráficas de datos genomáticos y proteómicos.

.

" Los bioinformáticos han tenido que invertir mucho tiempo en matemática de programación y algoritmos de estadística en un horario corto, tradicionalmente," dijo Kristen Amuzzini, gerente del área de biotecnologia, farmacéutica e industria médica de, The MathWorks. Y continua : "La combinación de los productos de MATLAB ® de hoy y el nuevo Bioinformatics Toolbox entrega las herramientas que ellos necesitan para analizar los datos grandes y como resultado da elementos que identifican los puntos donde potencialmente se necesita desarrollar un remedio rápida y eficazmente."

No dude en preguntarnos por el nuevo **Toolbox de Bioinformática**, es parte de una solución agil y efectiva que puede representarle mucho ahorro de tiempo y dinero, en biotecnologia, farmacéutica y medicina.

Si lo desea Usted puede abrir en www.mathworks.com el seminario

2. EJEMPLO DE ANALISIS DE SECUENCIAS DEL GENOMA HUMANO:

SEQSTATSDEMO Example of sequence statistics with MATLAB

This demonstration looks at some statistics about the DNA content of the human mitochondrial genome.

Contents

Introduction

Mitochondria are generally the major energy production center in eukaryotes. The Genome repository at the NCBI contains more interesting information about the human mitochondrial genome.

```
web(http://www.ncbi.nlm.nih.gov/)
```

The consensus sequence of the human mitochondria genome has accession number NC_001807. The whole GenBank entry is quite large and this example only uses the nucleotide sequence, so you can use the **getgenbank** function with the 'SequenceOnly' flag to read just the sequence information into the MATLAB workspace.

```
mitochondria = getgenbank('NC_001807','SequenceOnly',true);
```

If you don't have a live web connection, you can load the data from a MAT-file using the command

```
% load mitochondria     % <== Uncomment this if no internet connection
```

The MATLAB **whos** command gives information about the size of the sequence.

```
whos mitochondria
```

```
  Name               Size                    Bytes  Class

  mitochondria       1x16571                 33142  char array

Grand total is 16571 elements using 33142 bytes
```

You will use some of the sequence statistics function in the Bioinformatics Toolbox to look at various properties of this sequence. You can look at the composition of the nucleotides with the **ntdensity** function.

```
ntdensity(mitochondria)
```

Composition of the mitochondrial genome.

This shows that the genome is A-T rich. You can get more specific information with the **basecount** function.

```
    basecount(mitochondria)
ans =

    A: 5113
    C: 5192
    G: 2180
    T: 4086
```

These are on the 5'-3' strand. You can look at the reverse complement using the **seqrcomplement** function.

```
    basecount(seqrcomplement(mitochondria))
ans =

    A: 4086
    C: 2180
    G: 5192
    T: 5113
```

As expected, the base counts on the reverse complement strand are complementary to the counts on the 5'-3' strand.

You can use the chart option to **basecount** to display a pie chart of the distribution of the bases.

```
    figure
    basecount(mitochondria,'chart','pie');
```

Now look at the dimers in the sequence and display the information in a bar chart using **dimercount**.

```
    figure
    dimercount(mitochondria,'chart','bar')
ans =

    AA: 1594
    AC: 1495
    AG: 801
    AT: 1223
    CA: 1536
    CC: 1779
    CG: 439
    CT: 1438
    GA: 615
    GC: 716
    GG: 427
    GT: 421
    TA: 1368
    TC: 1202
    TG: 512
    TT: 1004
```

You can look at codons using **codoncount**. The function **dimercount** simply counts all adjacent nucleotides; however **codoncount** counts the codons on a particular reading frame. With no options, the function shows the codon counts on the first reading frame.

```
    codoncount(mitochondria)
AAA - 172     AAC - 157     AAG -  67     AAT - 123
ACA - 153     ACC - 163     ACG -  42     ACT - 130
AGA -  58     AGC -  90     AGG -  50     AGT -  43
ATA - 132     ATC - 103     ATG -  57     ATT -  96
CAA - 166     CAC - 167     CAG -  68     CAT - 135
CCA - 146     CCC - 215     CCG -  50     CCT - 182
CGA -  33     CGC -  60     CGG -  18     CGT -  20
CTA - 187     CTC - 126     CTG -  52     CTT -  98
GAA -  68     GAC -  62     GAG -  47     GAT -  39
GCA -  67     GCC -  87     GCG -  23     GCT -  61
GGA -  53     GGC -  61     GGG -  23     GGT -  25
GTA -  61     GTC -  49     GTG -  26     GTT -  36
TAA - 136     TAC - 127     TAG -  82     TAT - 107
TCA - 143     TCC - 126     TCG -  37     TCT - 103
TGA -  64     TGC -  35     TGG -  27     TGT -  25
TTA - 115     TTC - 113     TTG -  37     TTT -  99
```

Using a loop you can also look at all the other reading frames:

```
    for frame = 1:3
        figure
        subplot(2,1,1); codoncount(mitochondria,'frame',frame,'figure',true);
        title(sprintf('Codons for frame %d',frame));
        subplot(2,1,2);
    codoncount(mitochondria,'reverse',true,'frame',frame,'figure',true);
        title(sprintf('Codons for reverse frame %d',frame));
    end
```

Exploring the Open Reading Frames (ORFs)

In a nucleotide sequence an obvious thing to look for is if there are any open reading frames. The function **seqshoworfs** can be used to visualize ORFs in a sequence. Note: In the HTML tutorial only the first 7500 bases of the first reading frame are shown, however when running the demo you will be able to inspect the complete mitochondrial genome with the aid of the **Help Browser** .

```
    seqshoworfs(mitochondria);

Frame 1


000001    gatcacaggtctatcaccctattaaccactcacgggagctctccatgcatttggtattttcgtc
000065    tggggggtgtgcacgcgatagcattgcgagacgctggagccggagcaccctatgtcgcagtatc
000129    tgtctttgattcctgcctcattctattatttatcgcacctacgttcaatattacaggcgaacat
```

```
000193    acctactaaagtgtgttaattaattaatgcttgtaggacataataataacaattgaatgtctgc
000257    acagccgctttccacacagacatcataacaaaaaatttccaccaaaccccccctcccccgct
000321    tctggccacagcacttaaacacatctctgccaaaccccaaaaacaaagaaccctaacaccagcc
000385    taaccagatttcaaattttatctttaggcggtatgcacttttaacagtcacccccaactaaca
000449    cattattttcccctcccactcccatactactaatctcatcaatacaaccccgcccatcctacc
000513    cagcacacacacaccgctgctaaccccataccccgaaccaaccaaaccccaaagacacccccca
000577    cagtttatgtagcttacctcctcaaagcaatacactgaaaatgtttagacgggctcacatcacc
000641    ccataaacaaataggtttggtcctagcctttctattagctcttagtaagattacacatgcaagc
000705    atccccgttccagtgagttcaccctctaaatcaccacgatcaaaagggacaagcatcaagcacg
000769    cagcaatgcagctcaaaacgcttagcctagccacacccccacgggaaacagcagtgattaacct
000833    ttagcaataaacgaaagtttaactaagctatactaaccccagggttggtcaatttcgtgccagc
000897    caccgcggtcacacgattaacccaagtcaatagaagccggcgtaaagagtgttttagatcaccc
000961    cctccccaataaagctaaaactcacctgagttgtaaaaaactccagttgacacaaaatagacta
001025    cgaaagtggctttaacatatctgaacacacaatagctaagacccaaactgggattagatacccc
001089    actatgcttagccctaaacctcaacagttaaatcaacaaaactgctcgccagaacactacgagc
001153    cacagcttaaaactcaaaggacctggcggtgcttcatatccctctagaggagcctgttctgtaa
001217    tcgataaaccccgatcaacctcaccacctcttgctcagcctatataccgccatcttcagcaaac
001281    cctgatgaaggctacaaagtaagcgcaagtacccacgtaaagacgttaggtcaaggtgtagccc
001345    atgaggtggcaagaaatgggctacattttctaccccagaaaactacgatagccttatgaaact
001409    taagggtcgaaggtggatttagcagtaaactgagagtagagtgcttagttgaacagggccctga
001473    agcgcgtacacaccgcccgtcaccctcctcaagtatacttcaaaggacatttaactaaaacccc
001537    tacgcatttatatagaggagacaagtcgtaacatggtaagtgtactggaaagtgcacttggacg
001601    aaccagagtgtagcttaacacaaagcacccaacttacacttaggagatttcaacttaacttgac
001665    cgctctgagctaaacctagccccaaacccactccaccttactaccagacaaccttagccaaacc
001729    atttacccaaataaagtataggcgatagaaattgaaacctggcgcaatagatatagtaccgcaa
001793    gggaaagatgaaaaattataaccaagcataatatagcaaggactaacccctataccttctgcat
001857    aatgaattaactagaaataactttgcaaggagagccaaagctaagacccccgaaaccagacgag
001921    ctacctaagaacagctaaaagagcacacccgtctatgtagcaaaatagtgggaagatttatagg
001985    tagaggcgacaaacctaccgagcctggtgatagctggttgtccaagatagaatcttagttcaac
002049    tttaaatttgcccacagaaccctctaaatccccttgtaaatttaactgttagtccaaagaggaa
002113    cagctctttggacactaggaaaaaaccttgtagagagagtaaaaaatttaacacccatagtagg
002177    cctaaaagcagccaccaattaagaaagcgttcaagctcaacacccactacctaaaaaatcccaa
002241    acatataactgaactcctcacacccaattggaccaatctatcaccctatagaagaactaatgtt
002305    agtataagtaacatgaaaacattctcctccgcataagcctgcgtcagatcaaaacactgaactg
002369    acaattaacagcccaatatctacaatcaaccaacaagtcattattaccctcactgtcaacccaa
002433    cacaggcatgctcataaggaaaggttaaaaaaagtaaaaggaactcggcaaaccttaccccgcc
002497    tgtttaccaaaaacatcacctctagcatcaccagtattagaggcaccgcctgcccagtgacaca
002561    tgtttaacggccgcggtaccctaaccgtgcaaaggtagcataatcacttgttccttaaataggg
002625    acctgtatgaatggctccacgagggttcagctgtctcttacttttaaccagtgaaattgacctg
002689    cccgtgaagaggcgggcatgacacagcaagacgagaagaccctatggagctttaatttattaat
002753    gcaaacagtacctaacaaacccacaggtcctaaactaccaaacctgcattaaaaatttcggttg
002817    gggcgacctcggagcagaacccaacctccgagcagtacatgctaagacttcaccagtcaaagcg
002881    aactactatactcaattgatccaataacttgaccaacggaacaagttaccctagggataacagc
002945    gcaatcctattctagagtccatatcaacaatagggtttacgacctcgatgttggatcaggacat
003009    cccgatggtgcagccgctattaaaggttcgtttgttcaacgattaaagtcctacgtgatctgag
003073    ttcagaccggagtaatccaggtcggtttctatctacttcaaattcctccctgtacgaaaggaca
003137    agagaaataaggcctacttcacaaagcgccttcccccgtaaatgatatcatctcaacttagtat
003201    tatacccacacccacccaagaacagggtttgttaagatggcagagcccggtaatcgcataaaac
003265    ttaaaactttacagtcagaggttcaattcctcttcttaacaacatacccatggccaacctccta
003329    ctcctcattgtacccattctaatcgcaatggcattcctaatgcttaccgaacgaaaaattctag
003393    gctatatacaactacgcaaaggccccaacgttgtaggcccctacgggctactacaacccttcgc
003457    tgacgccataaaactcttcaccaaagagcccctaaaacccgccacatctaccatcaccctctac
003521    atcaccgccccgaccttagctctcaccatcgctcttctactatgaaccccccctccccatacccca
003585    acccccctggtcaacctcaacctaggcctcctatttattctagccacctctagcctagccgttta
003649    ctcaatcctctgatcagggtgagcatcaaactcaaactacgccctgatcggcgcactgcgagca
003713    gtagcccaaacaatctcatatgaagtcaccctagccatcattctactatcaacattactaataa
003777    gtggctcctttaacctctccacccttatcacaacacaagaacacctctgattactcctgccatc
003841    atgacccttggccataatatgatttatctccacactagcagagaccaaccgaaccccccttcgac
```

```
003905    cttgccgaaggggagtccgaactagtctcaggcttcaacatcgaatacgccgcaggccccttcg
003969    ccctattcttcatagccgaatacacaaacattattataataaacaccctcaccactacaatctt
004033    cctaggaacaacatatgacgcactctcccctgaactctacacaacatattttgtcaccaagacc
004097    ctacttctaacctccctgttcttatgaattcgaacagcatacccccgattccgctacgaccaac
004161    tcatacacctcctatgaaaaaacttcctaccactcaccctagcattacttatatgatatgtctc
004225    catacccattacaatctccagcattccccctcaaacctaagaaatatgtctgataaaagagtta
004289    ctttgatagagtaaataataggagcttaaaccccctttatttctaggactatgagaatcgaaccc
004353    atccctgagaatccaaaattctccgtgccacctatcacaccccatcctaaagtaaggtcagcta
004417    aataagctatcgggcccatacccccgaaaatgttggttataccttcccgtactaattaatcccc
004481    tggcccaacccgtcatctactctaccatctttgcaggcacactcatcacagcgctaagctcgca
004545    ctgatttttacctgagtaggcctagaaataaacatgctagcttttattccagttctaaccaaa
004609    aaaataaaccctcgttccacagaagctgccatcaagtatttcctcacgcaagcaaccgcatcca
004673    taatccttctaatagctatcctcttcaacaatatactctccggacaatgaaccataaccaatac
004737    taccaatcaatactcatcattaataatcataatggctatagcaataaaactaggaatagccccc
004801    tttcacttctgagtcccagaggttacccaaggcacccctctgacatccggcctgcttcttctca
004865    catgacaaaaactagcccccatctcaatcatataccaaatctctccctcactaaacgtaagcct
004929    tctcctcactctctcaatcttatccatcatagcaggcagttgaggtggattaaaccaaacccag
004993    ctacgcaaaatcttagcatactcctcaattacccacataggatgaataatagcagttctaccgt
005057    acaaccctaacataaccattcttaatttaactatttatattatcctaactactaccgcattcct
005121    actactcaacttaaactccagcaccacgaccctactactatctcgcacctgaaacaagctaaca
005185    tgactaacacccttaattccatccaccctcctctccctaggaggcctgcccccgctaaccggct
005249    ttttgcccaaatgggccattatcgaagaattcacaaaaaacaatagcctcatcatccccaccat
005313    catagccaccatcaccctccttaacctctacttctacctacgcctaatctactccacctcaatc
005377    acactactccccatatctaacaacgtaaaaataaaatgacagtttgaacatacaaaacccaccc
005441    cattcctccccacactcatcgcccttaccacgctactcctacctatctcccctttatactaat
005505    aatcttatagaaatttaggttaaatacagaccaagagccttcaaagccctcagtaagttgcaat
005569    acttaatttctgcaacagctaaggactgcaaaaccccactctgcatcaactgaacgcaaatcag
005633    ccactttaattaagctaagcccttactagaccaatgggacttaaacccacaaacacttagttaa
005697    cagctaagcaccctaatcaactggcttcaatctacttctcccgccgccgggaaaaaaggcggga
005761    gaagccccggcaggtttgaagctgcttcttcgaatttgcaattcaatatgaaaatcacctcgga
005825    gctggtaaaaagaggcctaacccctgtctttagatttacagtccaatgcttcactcagccattt
005889    tacctcacccccactgatgttcgccgaccgttgactattctctacaaaccacaaagacattgga
005953    acactatacctattattcggcgcatgagctggagtcctaggcacagctctaagcctccttattc
006017    gagccgagctgggccagccaggcaaccttctaggtaacgaccacatctacaacgttatcgtcac
006081    agcccatgcatttgtaataatcttcttcatagtaatacccatcataatcggaggctttggcaac
006145    tgactagttcccctaataatcggtgcccccgatatggcgtttccccgcataaacaacataagct
006209    tctgactcttacctccctctctcctactcctgctcgcatctgctatagtggaggccggagcagg
006273    aacaggttgaacagtctaccctcccttagcagggaactactcccaccctggagcctccgtagac
006337    ctaaccatcttctccttacacctagcaggtgtctcctctatcttaggggccatcaatttcatca
006401    caacaattatcaatataaaaccccctgccataacccaataccaaacgcccctcttcgtctgatc
006465    cgtcctaatcacagcagtcctacttctcctatctctcccagtcctagctgctggcatcactata
006529    ctactaacagaccgcaacctcaacaccaccttcttcgaccccgccggaggaggagaccccattc
006593    tataccaacacctattctgatttttcggtcaccctgaagtttatattcttatcctaccaggctt
006657    cggaataatctcccatattgtaacttactactccggaaaaaaagaaccatttggatacataggt
006721    atggtctgagctatgatatcaattggcttcctagggtttatcgtgtgagcacaccatatattta
006785    cagtaggaatagacgtagacacacgagcatatttcacctccgctaccataatcatcgctatccc
006849    caccggcgtcaaagtatttagctgactcgccacactccacggaagcaatatgaaatgatctgct
006913    gcagtgctctgagccctaggattcatctttcttttcaccgtaggtggcctgactggcattgtat
006977    tagcaaactcatcactagacatcgtactacacgacacgtactacgttgtagctcacttccacta
007041    tgtcctatcaataggagctgtatttgccatcataggaggcttcattcactgatttcccctattc
007105    tcaggctacaccctagaccaaacctacgccaaaatccatttcactatcatattcatcggcgtaa
007169    atctaactttcttcccacaacactttctcggcctatccggaatgccccgacgttactcggacta
007233    ccccgatgcatacaccacatgaaacatcctatcatctgtaggctcattcatttctctaacagca
007297    gtaatattaataatttcatgatttgagaagccttcgcttcgaagcgaaaagtcctaatagtag
007361    aagaaccctccataaacctggagtgactatatggatgccccccaccctaccacacattcgaaga
007425    acccgtatacataaaatctagacaaaaaaggaaggaatcgaaccccccaaagctggtttcaagc
007489    caaccccatggc
```

If you compare this output to the genes shown on the NCBI page there seem to be slightly fewer ORFs, and hence fewer genes, than expected. Vertebrate mitochondria do not use the Standard genetic code so some codons have different meaning in mitochondrial genomes. For more information about using different genetic codes in MATLAB see the help for the function **geneticcode**.

```
    help geneticcode

 GENETICCODE returns a structure containing mappings for the genetic code.

    MAP = GENETICCODE returns a structure containing mapping for the Standard
    genetic code.

    GENETICCODE(ID) returns a structure of the mapping for alternate genetic
    codes, where ID is either the transl_table ID from the NCBI Genetics web
    page (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c)
    or one of the following supported names. NAME can be truncated to the
    first two characters of the name.

    ID  Name

        1       Standard
        2       Vertebrate Mitochondrial
        3       Yeast Mitochondrial
        4       Mold, Protozoan, and Coelenterate Mitochondrial and
Mycoplasma/Spiroplasma
        5       Invertebrate Mitochondrial
        6       Ciliate, Dasycladacean and Hexamita Nuclear
        9       Echinoderm Mitochondrial
        10      Euplotid Nuclear
        11      Bacterial and Plant Plastid
        12      Alternative Yeast Nuclear
        13      Ascidian Mitochondrial
        14      Flatworm Mitochondrial
        15      Blepharisma Nuclear
        16      Chlorophycean Mitochondrial
        21      Trematode Mitochondrial
        22      Scenedesmus Obliquus Mitochondrial
        23      Thraustochytrium Mitochondrial

    Examples:

        moldcode = geneticcode(4);
        wormcode = geneticcode('Flatworm Mitochondrial');

    See also NT2AA, REVGENETICCODE.
```

The 'GeneticCode' option to the **seqshoworfs** function allows you to look at the ORFs again but this time with the vertebrate mitochondrial genetic code. Notice that there are now two much larger ORFs on the first reading frame: One starting at position 4471 and the other starting at 5905. These correspond to the ND2 (NADH dehydrogenase subunit 2) and COX1 (cytochrome c oxidase subunit I) genes.

```
    orfs = seqshoworfs(mitochondria,'GeneticCode','Vertebrate Mitochondrial',...
           'alternativestart',true)
orfs =
```

```
    Start: [1x26 double]
     Stop: [1x26 double]
```

Frame 1

```
000001    gatcacaggtctatcaccctattaaccactcacgggagctctccatgcatttggtattttcgtc
000065    tgggggtgtgcacgcgatagcattgcgagacgctggagccggagcaccctatgtcgcagtatc
000129    tgtctttgattcctgcctcattctattatttatcgcacctacgttcaatattacaggcgaacat
000193    acctactaaagtgtgttaattaattaatgcttgtaggacataataataacaattgaatgtctgc
000257    acagccgctttccacacagacatcataacaaaaaatttccaccaaaccccccctcccccgct
000321    tctggccacagcacttaaacacatctctgccaaaccccaaaaacaaagaaccctaacaccagcc
000385    taaccagatttcaaattttatctttaggcggtatgcacttttaacagtcacccccaactaaca
000449    cattattttcccctcccactcccatactactaatctcatcaatacaacccccgcccatcctacc
000513    cagcacacacacaccgctgctaaccccataccccgaaccaaccaaaccccaaagacaccccca
000577    cagtttatgtagcttacctcctcaaagcaatacactgaaaatgtttagacgggctcacatcacc
000641    ccataaacaaataggtttggtcctagcctttctattagctcttagtaagattacacatgcaagc
000705    atccccgttccagtgagttcaccctctaaatcaccacgatcaaaagggacaagcatcaagcacg
000769    cagcaatgcagctcaaaacgcttagcctagccacacccccacgggaaacagcagtgattaacct
000833    ttagcaataaacgaaagtttaactaagctatactaaccccagggttggtcaatttcgtgccagc
000897    caccgcggtcacacgattaacccaagtcaatagaagccggcgtaaagagtgttttagatcaccc
000961    cctccccaataaagctaaaactcacctgagttgtaaaaaactccagttgacacaaaatagacta
001025    cgaaagtggctttaacatatctgaacacacaatagctaagacccaaactgggattagataccc
001089    actatgcttagccctaaacctcaacagttaaatcaacaaaactgctcgccagaacactacgagc
001153    cacagcttaaaactcaaaggacctggcggtgcttcatatccctctagaggagcctgttctgtaa
001217    tcgataaaccccgatcaacctcaccacctcttgctcagcctatataccgccatcttcagcaaac
001281    cctgatgaaggctacaaagtaagcgcaagtacccacgtaaagacgttaggtcaaggtgtagccc
001345    atgaggtggcaagaaatgggctacattttctaccccagaaaactacgatagcccttatgaaact
001409    taagggtcgaaggtggatttagcagtaaactgagagtagagtgcttagttgaacagggccctga
001473    agcgcgtacacaccgcccgtcaccctcctcaagtatacttcaaaggacatttaactaaaacccc
001537    tacgcatttatatagaggagacaagtcgtaacatggtaagtgtactggaaagtgcacttggacg
001601    aaccagagtgtagcttaacacaaagcacccaacttacacttaggagatttcaacttaacttgac
001665    cgctctgagctaaacctagccccaaacccactccaccttactaccagacaaccttagccaaacc
001729    atttacccaaataaagtataggcgatagaaattgaaacctggcgcaatagatatagtaccgcaa
001793    gggaaagatgaaaaattataaccaagcataatatagcaaggactaacccctataccttctgcat
001857    aatgaattaactagaaataactttgcaaggagagccaaagctaagacccccgaaaccagacgag
001921    ctacctaagaacagctaaaagagcacacccgtctatgtagcaaaatagtgggaagatttatagg
001985    tagaggcgacaaacctaccgagcctggtgatagctggttgtccaagatagaatcttagttcaac
002049    tttaaatttgcccacagaaccctctaaatccccttgtaaatttaactgttagtccaaagaggaa
002113    cagctctttggacactaggaaaaaaccttgtagagagagtaaaaaatttaacacccatagtagg
002177    cctaaaagcagccaccaattaagaaagcgttcaagctcaacacccactacctaaaaaatcccaa
002241    acatataactgaactcctcacacccaattggaccaatctatcaccctatagaagaactaatgtt
002305    agtataagtaacatgaaaacattctcctccgcataagcctgcgtcagatcaaaacactgaactg
002369    acaattaacagcccaatatctacaatcaaccaacaagtcattattaccctcactgtcaacccaa
002433    cacaggcatgctcataaggaaaggttaaaaaaagtaaaaggaactcggcaaaccttaccccgcc
002497    tgtttaccaaaaacatcacctctagcatcaccagtattagaggcaccgcctgcccagtgacaca
002561    tgtttaacggccgcggtaccctaaccgtgcaaaggtagcataatcacttgttccttaaataggg
002625    acctgtatgaatggctccacgagggttcagctgtctcttacttttaaccagtgaaattgacctg
002689    cccgtgaagaggcgggcatgacacagcaagacgagaagaccctatggagctttaatttattaat
002753    gcaaacagtacctaacaaacccacaggtcctaaactaccaaacctgcattaaaaatttcggttg
002817    gggcgacctcggagcagaacccaacctccgagcagtacatgctaagacttcaccagtcaaagcg
002881    aactactatactcaattgatccaataacttgaccaacggaacaagttaccctagggataacagc
002945    gcaatcctattctagagtccatatcaacaatagggtttacgacctcgatgttggatcaggacat
003009    cccgatggtgcagccgctattaaaggttcgtttgttcaacgattaaagtcctacgtgatctgag
003073    ttcagaccggagtaatccaggtcggtttctatctacttcaaattcctccctgtacgaaaggaca
003137    agagaaataaggcctacttcacaaagcgccttcccccgtaaatgatatcatctcaacttagtat
```

```
003201    tatacccacacccacccaagaacagggtttgttaagatggcagagcccggtaatcgcataaaac
003265    ttaaaactttacagtcagaggttcaattcctcttcttaacaacatacccatggccaacctccta
003329    ctcctcattgtacccattctaatcgcaatggcattcctaatgcttaccgaacgaaaaattctag
003393    gctatatacaactacgcaaaggccccaacgttgtaggcccctacgggctactacaacccttcgc
003457    tgacgccataaaactcttcaccaaagagcccctaaaacccgccacatctaccatcaccctctac
003521    atcaccgccccgaccttagctctcaccatcgctcttctactatgaaccccctccccatacccca
003585    acccccctggtcaacctcaacctaggcctcctatttattctagccacctctagcctagccgttta
003649    ctcaatcctctgatcagggtgagcatcaaactcaaactacgccctgatcggcgcactgcgagca
003713    gtagcccaaacaatctcatatgaagtcaccctagccatcattctactatcaacattactaataa
003777    gtggctcctttaacctctccacccttatcacaacacaagaacacctctgattactcctgccatc
003841    atgacccttggccataatatgatttatctccacactagcagagaccaaccgaaccccttcgac
003905    cttgccgaaggggagtccgaactagtctcaggcttcaacatcgaatacgccgcaggcccttcg
003969    ccctattcttcatagccgaatacacaaacattattataataaacaccctcaccactacaatctt
004033    cctaggaacaacatatgacgcactctcccctgaactctacacaacatattttgtcaccaagacc
004097    ctacttctaacctcccctgttcttatgaattcgaacagcatacccccgattccgctacgaccaac
004161    tcatacacctcctatgaaaaaacttcctaccactcaccctagcattacttatatgatatgtctc
004225    catacccattacaatctccagcattccccctcaaacctaagaaatatgtctgataaaagagtta
004289    ctttgatagagtaaataataggagcttaaacccccttatttctaggactatgagaatcgaaccc
004353    atccctgagaatccaaaattctccgtgccacctatcacaccccatcctaaagtaaggtcagcta
004417    aataagctatcgggcccatacccgaaaatgttggttatacccttcccgtactaattaatcccc
004481    tggcccaacccgtcatctactctaccatctttgcaggcacactcatcacagcgctaagctcgca
004545    ctgatttttacctgagtaggcctagaaataaacatgctagcttttattccagttctaaccaaa
004609    aaaataaaccctcgttccacagaagctgccatcaagtatttcctcacgcaagcaaccgcatcca
004673    taatccttctaatagctatcctcttcaacaatatactctccggacaatgaaccataaccaatac
004737    taccaatcaatactcatcattaataatcataatggctatagcaataaaactaggaatagccccc
004801    tttcacttctgagtcccagaggttacccaaggcacccctctgacatccggcctgcttcttctca
004865    catgacaaaaactagccccatctcaatcatataccaaatctctccctcactaaacgtaagcct
004929    tctcctcactctctcaatcttatccatcatagcaggcagttgaggtggattaaaccaaacccag
004993    ctacgcaaaatcttagcatactcctcaattacccacataggatgaataatagcagttctaccgt
005057    acaaccctaacataaccattcttaatttaactatttatattatcctaactactaccgcattcct
005121    actactcaacttaaactccagcaccacgaccctactactatctcgcacctgaaacaagctaaca
005185    tgactaacacccttaattccatccaccctcctctccctaggaggcctgcccccgctaaccggct
005249    ttttgcccaaatgggccattatcgaagaattcacaaaaaacaatagcctcatcatccccaccat
005313    catagccaccatcaccctccttaacctctacttctacctacgcctaatctactccacctcaatc
005377    acactactccccatatctaacaacgtaaaaataaaatgacagtttgaacatacaaaacccaccc
005441    cattcctccccacactcatcgcccttaccacgctactcctacctatctcccctttatactaat
005505    aatcttatagaaatttaggttaaatacagaccaagagccttcaaagccctcagtaagttgcaat
005569    acttaatttctgcaacagctaaggactgcaaaaccccactctgcatcaactgaacgcaaatcag
005633    ccactttaattaagctaagcccttactagaccaatgggacttaaacccacaaacacttagttaa
005697    cagctaagcaccctaatcaactggcttcaatctacttctcccgccgccgggaaaaaaggcggga
005761    gaagccccggcaggtttgaagctgcttcttcgaatttgcaattcaatatgaaaatcacctcgga
005825    gctggtaaaaagaggcctaacccctgtctttagatttacagtccaatgcttcactcagccattt
005889    tacctcacccccactgatgttcgccgaccgttgactattctctacaaaccacaaagacattgga
005953    acactatacctattattcggcgcatgagctggagtcctaggcacagctctaagcctccttattc
006017    gagccgagctgggccagccaggcaaccttctaggtaacgaccacatctacaacgttatcgtcac
006081    agcccatgcatttgtaataatcttcttcatagtaatacccatcataatcggaggctttggcaac
006145    tgactagttcccctaataatcggtgcccccgatatggcgtttccccgcataaacaacataagct
006209    tctgactcttacctccctctctcctactcctgctcgcatctgctatagtggaggccggagcagg
006273    aacaggttgaacagtctaccctcccttagcagggaactactcccaccctggagcctccgtagac
006337    ctaaccatcttctccttacacctagcaggtgtctcctctatcttaggggccatcaatttcatca
006401    caacaattatcaatataaaaccccctgccataacccaataccaaacgcccctcttcgtctgatc
006465    cgtcctaatcacagcagtcctacttctcctatctctcccagtcctagctgctggcatcactata
006529    ctactaacagaccgcaacctcaacaccaccttcttcgaccccgccggaggaggagaccccattc
006593    tataccaacacctattctgatttttcggtcaccctgaagtttatattcttatcctaccaggctt
006657    cggaataatctcccatattgtaacttactactccggaaaaaaagaaccatttggatacataggt
006721    atggtctgagctatgatatcaattggcttcctagggtttatcgtgtgagcacaccatatattta
006785    cagtaggaatagacgtagacacacgagcatatttcacctccgctaccataatcatcgctatccc
006849    caccggcgtcaaagtatttagctgactcgccacactccacggaagcaatatgaaatgatctgct
```

```
006913    gcagtgctctgagccctaggattcatctttcttttcaccgtaggtggcctgactggcattgtat
006977    tagcaaactcatcactagacatcgtactacacgacacgtactacgttgtagctcacttccacta
007041    tgtcctatcaataggagctgtatttgccatcataggaggcttcattcactgatttcccctattc
007105    tcaggctacaccctagaccaaacctacgccaaaatccatttcactatcatattcatcggcgtaa
007169    atctaactttcttcccacaacactttctcggcctatccggaatgccccgacgttactcggacta
007233    ccccgatgcatacaccacatgaaacatcctatcatctgtaggctcattcatttctctaacagca
007297    gtaatattaataattttcatgatttgagaagccttcgcttcgaagcgaaaagtcctaatagtag
007361    aagaaccctccataaacctggagtgactatatggatgccccccaccctaccacacattcgaaga
007425    acccgtatacataaaatctagacaaaaaaggaaggaatcgaacccccccaaagctggtttcaagc
007489    caaccccatggc
```

Extracting and analyzing the ND2 protein

The ORF of interest starts at position 4471, the following commands can be used to find the corresponding stop codon:

```
    ND2Start = 4471;
    startIndex = find(orfs(1).Start == ND2Start)
    ND2Stop = orfs(1).Stop(startIndex)
startIndex =

    24


ND2Stop =

        5512
```

Once the positions are known, MATLAB indexing can be used to extract the region of interest.

```
    ND2Seq = mitochondria(ND2Start:ND2Stop);
```

If you look at the **codoncount** for this gene we see a lot of CTA and ATC codons.

```
    codoncount(ND2Seq)
AAA - 10     AAC - 14     AAG -  2     AAT -  6
ACA - 11     ACC - 24     ACG -  3     ACT -  5
AGA -  0     AGC -  4     AGG -  0     AGT -  1
ATA - 22     ATC - 24     ATG -  2     ATT -  8
CAA -  8     CAC -  3     CAG -  2     CAT -  1
CCA -  4     CCC - 12     CCG -  2     CCT -  5
CGA -  0     CGC -  3     CGG -  0     CGT -  1
CTA - 26     CTC - 18     CTG -  4     CTT -  7
GAA -  5     GAC -  0     GAG -  1     GAT -  0
GCA -  8     GCC -  7     GCG -  1     GCT -  4
GGA -  5     GGC -  7     GGG -  0     GGT -  1
GTA -  3     GTC -  2     GTG -  0     GTT -  3
TAA -  0     TAC -  8     TAG -  0     TAT -  2
TCA -  7     TCC - 11     TCG -  1     TCT -  4
TGA - 10     TGC -  0     TGG -  1     TGT -  0
TTA -  8     TTC -  7     TTG -  1     TTT -  8
```

For those of you who have not memorized the genetic code you can easily check what amino acids these codons get translated into using the **nt2aa** and **aminolookup** functions.

```
    aminolookup('letter',nt2aa('CTA'))
    aminolookup('letter',nt2aa('ATC'))
ans =

Leu     leucine


ans =

Ile     isoleucine
```

The **nt2aa** function converts the nucleotide sequence to the corresponding amino acid sequence. Again the 'GeneticCode' option must be used to specify the vertebrate mitochondrial genetic code.

```
    ND2 = nt2aa(ND2Seq,'GeneticCode','Vertebrate Mitochondria');
```

You can get a more complete picture of the amino acid content with **aacount**.

```
    figure
    aacount(ND2,'chart','bar')
ans =

    A: 20
    R: 4
    N: 20
    D: 0
    C: 0
    Q: 10
    E: 6
    G: 13
    H: 4
    I: 31
    L: 64
    K: 12
    M: 25
    F: 15
    P: 23
    S: 28
    T: 43
    W: 11
    Y: 10
    V: 8
```

Notice the high leucine, threonine and isoleucine content and also the lack of cysteine or aspartic acid.

You can use the **atomiccomp** and **molweight** functions to find out more about the ND2 protein.

```
    atomiccomp(ND2)
    molweight(ND2)
ans =

    C: 1818
    H: 2882
    N: 420
    O: 471
    S: 25


ans =

  3.8960e+004
```

For further investigation of the properties of the ND2 protein, try using **proteinplot**. This is a graphical user interface (GUI) that allows you to easily create plots of various properties, such as hydrophobicity, of a protein sequence. Click on the "Help" menu in the GUI for more information on how to use the tool.

```
    proteinplot(ND2)
```

3. ANÁLISIS FILOGENÉTICO

- Proceso que usamos para determinar la relación evolutiva entre organismos.
- Estos resultados pueden dibujarse en un diagrama jerárquico llamado filograma (árbol filogenético)

EJEMPLO: CONSTRUCCIÓN DE UN ARBOL FILOGENÉTICO

- Usamos datos de secuencias mitocondriales (D-loop), creamos un árbol filogenético para una familia de primates.
- A partir de secuencias mitocondriales DNA (mtDNA) para la familia HOMINIDAE.

3.1. BÚSQUEDA DE DATOS FILOGENÉTICOS EN NCBI

PROCEDIMIENTO

- Use the MATLAB Help browser to search for data on the Web. In the MATLAB Command Window, type web('http://www.ncbi.nlm.nih.gov')
- Buscar el sitio web NCBI para información
- Seleccionar el link taxonomy para la familia HOMINIDAE

3.2. CREACIÓN DE UN ARBOL FILOGENÉTICO PARA CINCO ESPECIES

- Crear una estructura MATLAB con información acerca de las secuencias.
- Se emplean códigos de acceso para las secuencias mitocondriales D-loop aisladas de especies homínidas diferentes.

```
data = {'German_Neanderthal'      'AF011222';
        'Russian_Neanderthal'     'AF254446';
        'European_Human'          'X90314'  ;
```

```
'Mountain_Gorilla_Rwanda' 'AF089820';
'Chimp_Troglodytes'       'AF176766'; };
```

- Obtener secuencias de datos de la base de datos GenBank y copiarlas en MATLAB

```
for ind = 1:5
   seqs(ind).Header   = data{ind,1};
   seqs(ind).Sequence = getgenbank(data{ind,2},
                                   'sequenceonly', true);
end
```

- Se calculan las distancias de parejas discretas y se crea un objeto phytree

```
distances = seqpdist(seqs,'Method','Jukes-Cantor','Alphabet','DNA');
            tree = seqlinkage(distances,'UPGMA',seqs)
```

- MATLAB dibuja un árbol filogenético

```
h = plot(tree,'orient','bottom');
ylabel('Evolutionary distance')
set(h.terminalNodeLabels,'Rotation',-45)
```

3.3. CREACION DE UN ARBOL FILOGENÉTICO PARA DOCE ESPECIES

- Se adicionan más secuencias a la estructura MATLAB

```
data2 = {'Puti_Orangutan'          'AF451972';
         'Jari_Orangutan'          'AF451964';
         'Western_Lowland_Gorilla' 'AY079510';
         'Eastern_Lowland_Gorilla' 'AF050738';
         'Chimp_Schweinfurthii'    'AF176722';
         'Chimp_Vellerosus'        'AF315498';
```

```
'Chimp_Verus'          'AF176731';
                    };
```

- Obtener las secuencias adicionales de datos adicionales de la base de datos GenBank.

```
for ind = 1:7
    seqs(ind+5).Header   = data2{ind,1};
seqs(ind+5).Sequence = getgenbank(data2{ind,2},
                       'sequenceonly', true);
end
```

- Se calculan las distancias pares iguales y el apareamiento jerárquico

```
distances = seqpdist(seqs,'Method','Jukes-Cantor','Alpha','DNA');
tree = seqlinkage(distances,'UPGMA',seqs);
```

- MATLAB dibuja un árbol filogenético

```
h = plot(tree,'orient','bottom');
ylabel('Evolutionary distance')
set(h.terminalNodeLabels,'Rotation',-45)
```

4. Referencias:

4.1. BioinformaticsToolboxForUsewithMATLAB®, User Guide, V. 21.1, The MathWorks Inc. 2005.
4.2. MATLAB 7.1, Release 14 Service Pack 3, The MathWorks Inc. 2005.
4.3. Bioinformatics Toolbox 2.1.1. The MathWorks Inc. , Septiembre 2005,