# BIOWORD DEVELOPER'S MANUAL

## Why Microsoft Word and VBA?

Microsoft Word is a very commonly used program among biologists with a built-in programming language (VBA) that is readily accessible. By using macros that are built into a Microsoft Word document, only one file is required for installation (the .dotm file). The Ribbon is created using XML and edited using Microsoft Word's Custom UI Editor (downloadable from http://openxmldeveloper.org/articles/customuieditor.aspx)

Word 2007+ documents are saved in Open XML format, which allows us to save the options necessary for sequence operations in the Document file as a Custom XML Part.

## CLASSES

### *SEQUENCE*

The Sequence class is for DNA, Protein or RNA sequences. This class determines what a sequence is by counting up the number of A's, C's, G's and T's/U's and comparing them to the user-designated threshold for DNA. The Sequence class is responsible for filtering unwanted characters from the sequences.

An empty Sequence must first be created and then have its information set because VBA does not allow instance variables.

The Sequence's sequence is stored as a String variable, which can be accessed as raw text by calling its `convertToRaw` method.

### *COLSEQUENCES*

The ColSequences class holds a Collection of Sequences. This class accesses the individual Sequences' methods and sends the results back to the RibbonControl module.

ColSequences can be initialized in two ways—first as the sequences that are selected in the document, and second as an empty collection to which new sequences may be added (to use this option, when initializing the ColSequences object, set the `refColl` parameter to `True`)

The ColSequences class can be used to generate PSFMs for its collection of sequences, making it useful for motifs.

Most of the code and inter-class method calling for the functions is found here.

## GCODE

The GCode class contains information relating to codons and corresponding amino acids. It is chiefly used in translating and reverse translating, as well as any function that requires a codon usage table. It has 9 genetic codes hard coded, though only one is initialized for a given instance of the class.

## ALIGNMENTCELL

The AlignmentCell class is used in the pair-wise alignment methods. It consists of 4 fields: the score of a cell in the alignment matrix and then whether it was reached from a vertical move, a horizontal move and/or a diagonal move.

## SCOREMATRIX

The Score Matrix class is used in scoring matches and mismatches when needed in pattern matching and alignments. ScoreMatrix contains the BLOSUM62 matrix, which is hardcoded as a Collection, as well as another Collection that associates each IUB character with its possibilities.

To add a new scoring matrix:

1) In the ScoreMatrix class, go to the `fillMatrix` Sub, add:

```
ElseIf(protMethod = x) Then

      Call scoreMatrix.add(Array(...), “<base>”)

      ...

End If
```

The array should be the score of matching <base> with each amino acid. The array must be in this order:

```
A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
```

2) In the AdvOpt UserForm, go to the `UserForm_Initialize` Sub. Locate this line:

```
scoringMethod.List = Array("BLOSUM62", "PAM250", "PAM120",
"BLOSUM50", "GONNET")
```

Add the name of your new matrix to the end of the list.

### *EVENTHANDLER*

The EventHandler class is used to save the BioWord.dotm file automatically when Word is closed without Word prompting for changes to be saved.

## MODULES

### *RIBBONCONTROL*

The RibbonControl module is responsible for communicating between the ColSequences method and the document. It contains the methods that the buttons of the ribbon are linked to and creates the initial ColSequences object. It also calls the appropriate ColSequences methods based on which button was pressed. The RibbonControl receives all results from the ColSequences object (in the form of a Collection or a String) and calls the `enterToDocument` or the `enterHighlight` methods of the Resources module to print the results.

### *RESOURCES*

The Resources module contains a variety of generic functions that can also be performed on non-sequences. For example, it contains a method that will sort a Collection, remove duplicates for a collection, and print Strings, Collections and Tables to the document.

### *XMLHANDLER*

The XMLHandler module allows modifications of the options XML file, including changing the names and values of the options 'nodes', adding and removing 'nodes' and adding a new XML options file

## RIBBON OPTIONS XML FILE

NOTE: COMMENTS ARE **NOT** INCLUDED IN ORIGINAL FILE

```
<?xml version="1.0" standalone="yes" ?>
<options>
      <outputMode>1</outputMode>              // used to determine which
                                              format the results should be
                                              printed in
```

```
                                        0 = Raw
                                        1 = FASTA
                                        2 = GenBank

<outputLoc>0</outputLoc>                // used to determine where the
                                        results should be printed

                                        0 = Below Selection
                                        1 = New Document
                                        2 = Save to Clipboard
                                        3 = Replace Selection

<isDNA>70</isDNA>                       // the minimum percentage of
                                        A’s, C’s, T’s and G’s that is
                                        required for a sequence to be
                                        considered DNA

<revTran>0</revTran>                    // used to determine which
                                        form of reverse translation is
                                        to be used

                                        0 = Uniform
                                        1 = IUB
                                        2 = Best Codon
                                        3 = Random Best Codon

<offset>0</offset>                      // used to determine the frame
                                        of translation

                                        0, 1, or 2

<strict>1</strict>                      // used to determine if IUB
                                        characters are allowed in a
                                        sequence

                                        0 = NO IUB characters
                                        1 = IUB characters allowed

<gcode>0</gcode>                        // used to determine which
                                        genetic code should be used by
                                        default

                                        0 = Standard
                                        1 = Vertebrate Mitochondrial
                                        2 = Yeast Mitochondrial
                                        3 = Mold Mitochondrial
                                        4 = Invertebrate Mitochondrial
                                        5 = Ciliate Nuclear
                                        6 = Echinoderm Mitochondrial
                                        7 = Euplotid Mitochondrial
                                        8 = Bacterial Plasmid

<usageFormat>0</usageFormat>            // used to determine what
```

```
                                        format the codon usage table
                                        should be printed in

                                        0 = Table format
                                        1 = Whitespace format

<usageOffset>0</usageOffset>            // used to determine the frame
                                        of codon usage table

                                        0, 1, or 2

<winLen>10</winLen>                     // used to determine the
                                        length of the window when %
                                        window GC is calculated

<stepSize>1</stepSize>                  // used to determine how much
                                        to move the sliding window
                                        when % window GC is calculated

<nGram>2</nGram>                        // the size of an N-Gram

<nGramOpt>0</nGramOpt>                  // used to determine whether
                                        all N-Grams will be printed or
                                        only the ones found in the
                                        sequence

                                        0 = all
                                        1 = only found in the sequence

<nGramRev>0</nGramRev>                  // used to determine whether
                                        the reverse complement of the
                                        sequence will be included in
                                        the N-Gram results

                                        0 = do not include rev. comp.
                                        1 = include rev. comp.

<monoOpt>0</monoOpt>                    // used to determine if the 5’
                                        monophosphate will be included
                                        when calculating the molecular
                                        weight of DNA sequences
                                        0 = do not include
                                        1 = include

<triOpt>0</triOpt>                      // used to determine if the 5’
                                        triphosphate will be included
                                        when calculating the molecular
                                        weight of RNA sequences
                                        0 = do not include
                                        1 = include
```

```
<dblStrand>0</dblStrand>                 // used to determine if DNA
                                         sequences will be considered
                                         double-stranded when
                                         calculating the molecular
                                         weight

                                         0 = considered single-stranded
                                         1 = considered double-stranded

<minLength>10</minLength>                // the minimum length (in
                                         codons) that an ORF has to be
                                         to be reported

<useCAI>0</useCAI>                       // used to determine whether
                                         ORFs will be prioritized based
                                         on length or a combination of
                                         length and CAI

                                         0 = length only
                                         1 = CAI + length

<maxMismatch>0</maxMismatch>             // the maximum threshold for
                                         mismatches in Substring and
                                         Dyad Searches

<revComp>0</revComp>                     // used to determine the
                                         second dyad motif will be the
                                         reverse complement or
                                         duplicate of the original dyad
                                         motif

                                         0 = Mirror motif (rev. comp.)
                                         1 = Duplicate

<multiplyFactor>1.5</multiplyFactor>     // the factor times the
                                         information content (IC) that
                                         determines the minimum score a
                                         sequence must meet to be be
                                         reported

<endGapRng>0</endGapRng>                 // the maximum size to a gap
                                         in Substring with Gap and Dyad
                                         Pattern Search

<begGapRng>0</begGapRng>                 // the minimum size to a gap
                                         in Substring with Gap and Dyad
                                         Pattern Search

<genomeGC>50</genomeGC>                  //  the percentage of G’s and
                                         C’s in a genome; used for
                                         Gibbs Sampling, Ri Sequence
                                         Scoring, I Sequence Scoring
```

```
                                        and Greedy Sampling

<spacerLen>1</spacerLen>                // the base length for spacer
                                        when running Dyad Motif
                                        Discovery

<dyadLen>5</dyadLen>                    // the base length for dyads
                                        in Dyad Motif Discovery

<spacerInc>1</spacerInc>                // the range for spacer length
                                        when running Dyad Motif
                                        Discovery

<dyadInc>1</dyadInc>                    // the range for dyad length
                                        when running Dyad Motif
                                        Discovery

<palindrome>0</palindrome>              // used to determine if the
                                        second dyad will be the a
                                        duplicate motif or a
                                        palindrome in Dyad Motif
                                        Discovery

                                        0 = Palindrome
                                        1 = Duplicate

 <numIts>100</numIts>                   // the number of iterations to
                                        run the Gibbs Sampling/Greedy
                                        Search to find the motif with
                                        the highest IC

<gibbsWinLen>10</gibbsWinLen>           // the size of the motif to be
                                        found in Gibbs Sampling/Greedy
                                        Search

<alignMatch>2</alignMatch>              // the match score in pair-
                                        wise alignments

<alignMis>0</alignMis>                  // the mismatch score in pair-
                                        wise alignments

<alignGOP>-2</alignGOP>                 // the gap opening penalty in
                                        pair-wise alignments

 <alignGEP>-1</alignGEP>                // the gap extension penalty
                                        in pair-wise alignments

<maxAlign>1</maxAlign>                  // the maximum number of
                                        results of a pair-wise
                                        alignment
```

```
<at>0</at>                                  // used to fill in the scoring
                                            matrix for DNA in Pair-wise
<ac>0</ac>                                  Alignment

<ag>0</ag>                                  Ex. <at> would represent the
                                            mismatch score of matching an
<aa>2</aa>                                  A with a T or a T with an A

<tt>2</tt>

<tc>0</tc>

<tg>0</tg>

<cg>0</cg>

<cc>2</cc>

<gg>2</gg>

<useMatrix>0</useMatrix>                    // used to determine whether
                                            the match/mismatch or matrix
                                            option will be used in scoring
                                            DNA for a Pair-wise Alignment

                                            0 = Use match/mismatch scoring
                                            1 = Use scoring matrix

<useRi>0</useRi>                            // used to determine the
                                            scoring function for site
                                            searches

                                            0 = use Ri Sequence
                                            1 = use I sequence

<useRSeq>0</useRSeq>                        // used to determine the
                                            method to calculate
                                            Information Content (IC)

                                            0 = R Sequence
                                            1 = Relative Entropy

<logoIUB>0</logoIUB>                        // used to determine if IUB
                                            characters are used to replace
                                            equally probable bases in
                                            consensus sequences

                                            0 = do not use IUB characters
                                            1 = use IUB characters

<maxResults>4</maxResults>                  // used to determine how many
                                            results are printed in the
                                            site search method

<pseudocount>0</pseudocount>                // used to determine the
                                            method to calculate
```

```
                                         pseudocounts

                                         0 = LaPlace's method
                                         1 = 10^-50

  <protScore>0</protScore>               // used to determine the
                                         scoring matrix for protein
                                         sequences

                                         0 = BLOSUM62


  <wrapFASTA>1</wrapFASTA>               // used to determine whether
                                         FASTA sequences should wrap to
                                         90 characters

                                         0 = do not wrap

                                         1 = do wrap


  <useBLOSUM>0</useBLOSUM>               // used to determine whether
                                         to use a scoring matrix (at
                                         the time of this writing only
                                         BLOSUM62 was available)
                                         instead of string match


                                         0 = use scoring matrix

                                         1 = use mismatch penalty


  <BLOSUMThres>19</BLOSUSMThres>         // the minimum threshold for a
                                         score in Substring and Dyad
                                         Searches (for protein
                                         sequences)


  <bit>2</bit>                           // used to scale the height of
                                         the reference vertical bar in
                                         sequence logos

<precision>3</precision>                 // how many decimal places to
```

```
                                             print numerical answers to

   <includePSFM>0</includePSFM>              // whether or not to display a
                                             PSFM with a consensus logo

                                             0 = do not display PSFM

                                             1 = display PSFM

   <protGenome>0</protGenome>                // background frequencies for
                                             amino acids

                                              0 = uniform

                                              1 = BLOSUM62 frequencies

   <genGraph>1</genGraph>                    // whether to generate a graph
                                             for %GC Window

                                              0 = no graph (table output)

                                              1 = graph

   <typePSFM>1</typePSFM>                    // PSFM format type for
                                             consensus logos

                                              0 = table format

                                              1 = Jaspar matrix format

   <epsilon>0.05</epsilon>                   // used as a buffer to
                                             determine if two bases are
                                             similarly frequent in
                                             consensus logos

   <orfRevComp>1</orfRevComp>                // whether to search reverse
                                             complement of sequence for
                                             ORFs

                                              0 = do not search

                                              1 = search reverse complement

</options>
```