BIOWORD: USER MANUAL

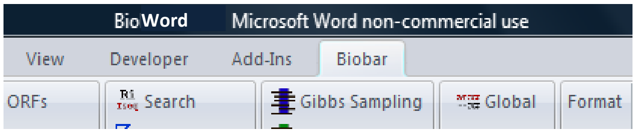
INSTALLATION:

AUTOMATIC INSTALLATION:

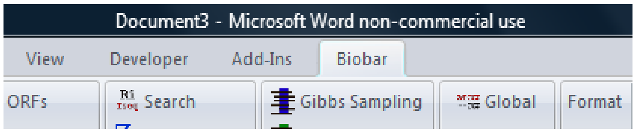
1) Open BioWord.dotm by **right-clicking the file icon and clicking "Open."** Alternatively, you can open the file from within Microsoft Word.
    - Do NOT double-click the file icon. BioWord is a template, which will by default open behind a new document.

Make sure '**BioWord**' is printed at the top of Microsoft Word—NOT '**Document**'.

CORRECT



INCORRECT



2) Double-click the "Click to Install BioWord" button in the document.
3) Follow the Installation prompts.

NOTE: If Word is installed in a non-default location, installation must be **manual.**

MANUAL INSTALLATION

The BioWord.dotm file needs to be placed in the Microsoft Word STARTUP folder so that the toolbar is available for every document. The path to this folder is listed below:

For Windows XP:

C:\Documents and Settings\<Username>\Application Data\Microsoft\Word\Startup

For Windows Vista/7:

C:\Users\<Username>\AppData\Roaming\Microsoft\Word\STARTUP

MACRO SECURITY

In order to use BioWord, you may need to change your macro security settings

1) Click the **circular Word button** in the upper left-hand corner
2) Go to **Word Options** at the bottom right of the menu

3) Click the **Trust Center** on the left hand menu
4) Click the button marked **Trust Center Settings**
5) Make sure **"Enable all macros…"** is selected (the last option)

SELECTION:

BioWord can recognize sequences entered in **raw**, **GenBank**, and **FASTA** format.

- Raw and GenBank sequences must be separated by **two enters.**
- If any sequences are entered in FASTA format (with a **>** denoting the header), the **>** will be used to separate sequences rather than two enters.

To perform an operation, simply **select the sequence(s)** you are interested in.

There are DNA/RNA-specific operations, protein-specific operations, and operations that can be performed on any type of sequence. *BioWord will only perform an operation on appropriate selected sequence(s)*. For example, if a DNA sequence and protein sequence are both selected for a DNA/RNA-specific operation, the operation will be performed only on the DNA sequence.

GLOBAL OPTIONS

BASIC OPTIONS

Default Output

Raw – your results will be displayed in the raw format

FASTA – your results will be displayed in the FASTA format (if a header was not provided, a numbered default one will be created)

GenBank – your results will be displayed in the GenBank format

Default Output Location

Append to Document – your results will be printed beneath the selection

Create New Document – your results will be printed on a new Document

Save to Clipboard – your results will be saved to the Clipboard; use Paste (*Ctrl + V*) to retrieve them (note that this option limits formatting)

Replace Selection – your results will overwrite the selection

ADVANCED OPTIONS (POP-UP)

%ATCG

This is the value used to determine what composition of bases defines a DNA/RNA sequence (for RNA, `T` is replaced by `U`). It should be entered as an integer percentage (i.e., 70 rather than .70).

If the “Allow IUB characters” option is selected, the sequence will be considered DNA/RNA irrespective of %ATCG if no non-IUB characters are found in the sequence.

Precision

All numerical results will be printed to this number of decimal places

Allow IUB characters

If this option is selected, the following characters will be allowed in DNA/RNA sequences:

`A, T/U, C, G, N(ATCG), V(GAC), B(GTC), H(ATC), D(GAT),`
`K(GT), S(GC), W(AT), M(AC), Y(CT), R(AG)`

All other characters will be filtered out.

If this option is not selected, only `A, C, T/U, G` will be allowed in DNA/RNA sequences. All other characters will be filtered out.

Ex) If `AGGATCGAMMA` were selected, `AGGATCGAA` would be processed.

Wrap FASTA Text

If this option is selected, FASTA sequences will be wrapped every 90 characters when printed

Pseudocounts

What are pseudocounts?

A pseudocount is a number that can be added to a quantity that is known to be non-zero. In a sequence motif for example, a zero frequency of a base at a certain position could simply be due to too small a dataset. A pseudocount accounts for and corrects this possibility.

How are they calculated?

BioWord offers two methods of calculating pseudocounts. Below are the formulas used for each:

LaPlace’s Rule: $\frac{obs + 2/M}{N + 2}$ where M is the number of alphabet size and N is the number of sequences

10^{-50}: Use this option to avoid taking the log of 0

Where are they used?

The pseudocounts are used for Position Specific Frequency Matrix (PSFM) calculations. A PSFM is generated for the following operations:

- Search
- Dyad Pattern Search
- Consensus Logo
- Gibbs Sampling
- Greedy Sampling

Information Content

What is Information Content (IC)?

Information Content is a value that quantifies the difference between the position-specific frequencies of a motif and a uniform or background distribution.

How is it calculated?

BioWord offers two method of calculating IC: Shannon Entropy (also called $R_{sequence}$) and Relative Entropy. Shannon Entropy concerns uncertainty, or how predictable a motif is given the expected distribution. Relative Entropy puts more emphasis on the background distribution: comparing the background entropy with the entropy of the motif. Below are the formulas used to calculate each:

$R_{sequence}$ (Shannon Entropy)

$$\sum_{i=1}^{N}\left(-\sum_{s\in\Omega} f_s \log_2(f_s) - -\sum_{s\in\Omega} p_{s,i}\log_2(p_{s,i})\right)$$

where f_s is the expected frequency of base *s* in the genome and $p_{s,i}$ is the frequency of base *s* in position *i* in the motif

Relative Entropy

$$\sum_{i=1}^{N}\left(-\sum_{s\in\Omega} p_{s,i}\log_2(f_s) - (-p_{s,i}\log_2(p_{s,i}))\right)$$

where f_s is the expected frequency of base *s* in the genome and $p_{s,i}$ is the frequency of base *s* in position *i* in the motif

Where is it used?

The calculated IC values are used for the following operations:

- Search
- Dyad Pattern Search
- Consensus Logo
- Gibbs Sampling
- Greedy Sampling

Search Scoring Function

What is a scoring function?

A scoring function can be used to search for a particular site in a sequence when given a sequence motif (the query) and some background genomic information. Higher scoring results are better matches.

How is it calculated?

BioWord offers two scoring functions: R_i sequence and $I_{sequence}$. R_i sequence is derived from the information content equation $R_{sequence}$, and $I_{sequence}$ is derived from the Relative Entropy equation. Below are the formulas used to calculate each:

R_i Sequence

$$\sum_{i=1}^{N}\left(-\sum_{s\in\Omega} f_s \log_2(f_s) - -\log_2(p_{s,i})\right)$$

where f_s is the expected frequency of base *s* in the genome and $p_{s,i}$ is the frequency of base *s* in position *i* in the motif

$I_{sequence}$

$$\sum_{i=1}^{N}\sum_{s\in\Omega} p_{s,i}\left(\log_2 p_{s,i} - \log_2 f_s\right)$$

where f_s is the expected frequency of base *s* in the genome and $p_{s,i}$ is the frequency of base *s* in position *i* in the motif

Where is it used?

The scores calculated from these functions are used for the following operations:

- Search
- Dyad Pattern Search
- Gibbs Sampling
- Greedy Sampling

Protein Scoring

Scoring Matrix

BLOSUM 62

PAM 250

PAM 120

BLOSUM 50

GONNET

The selected scoring matrix will be used primarily for alignments of protein sequences. However, if the check box "**use this instead of mismatches**" is selected, the selected scoring matrix will be used to score results for the following operations:

- Substring Search
- Gapped Substring Search
- Dyad Motif Discovery

If the checkbox is not selected, your results will be scored simply by the number of mismatches.

Protein Background Frequencies

Uniform-a uniform distribution; yields background entropy of $\log_2(20)$ = 4.32 bits

BLOSUM62 Frequencies-the distribution of amino acids used to create the BLOSUM62 matrix. Data can be found at http://selab.janelia.org/publications/Eddy-ATG2/lambda.c

TOOLS:

ADD COMMENTS

This function allows the user to add notes about a sequence or site. Select the text to which would like the comment to refer, and then click the "Add Comment" button on the Ribbon.

Comments can be removed by selecting the sequence and clicking "Remove Formatting" on the Ribbon, or by right-clicking on the comment bubble and selecting "Delete Comment"

REMOVE FORMATTING

Several of BioWord's search operations can highlight matching sites in the sequence. To remove any highlighting or text coloration from a sequence easily, select the sequence(s) and then click the "Remove Formatting" button on the Ribbon. As mentioned above, the "Remove Formatting" operation will also remove any comments that have been entered.

BIGGER/SMALLER

These text-size manipulation buttons are useful when generating Consensus Logos, as when a logo is originally printed, the text can be quite small. To increase the size of the logo while maintaining proportions, select the logo and click the "Bigger" button on the Ribbon until the size is acceptable. Similarly, the "Smaller" button can be used to shrink the text of a logo without disrupting the proportions.

MANIPULATION:

RAW

The **raw** format consists of the genetic sequence in plain text, with multiple sequences separated with two enters. By clicking the “Raw” button on the Ribbon, any selected sequences will be printed in the raw format. An example of a sequence in the raw format is shown below:

```
ATGATTTACCTGAAGTCCTTACTCAATGTTATTGATAATAGCGGGGCCCAGGTTGTCGAGTGTATCAAGGTCCTGCGG
CATAAGCCGAAGTCCTGTGCTCAGATTGGTGATCGTATTACCTGTGTCGTTAAGCAGGCGCGCCCCTTACAGCAGGAG
CTCACCGGTCAGTCGTCCACCAATCGTGTCAAGCGTCGCGATATCTGTCAGGCCGTCGTTGTCAGAACCCGCGCTCCG
CTTAAGCGCAAGGATGGTAGCGTCGTGAGGTTTGATGATAATGCCTGTGTCCTCATCAATAAGAATGGCGAGCCCCTC
```

FASTA

The **FASTA** format precedes each genetic sequence with a header. A header begins with the > character, and is separated from the sequence itself with a single enter. There is no restriction as to the characters that may be included in a header. In BioWord, a FASTA sequence is defined as all characters from the new-line after the header until the next > character, denoting another FASTA sequence. An example of a sequence in FASTA format is shown below:

```
>gi|17544719:1391233-1391883 Ralstonia solanacearum GMI1000, complete genome

TTATGGAGCGGCTGGCCGGATCAGGCCGACCGCCAGTCCTTCCAACTGGAACTCGTCGCGGTCGAGATCGACGTGGATG
GGTTCGAAATCAGGGTTCTCGGCAATCAGCTCGACCTGCCGGCCTTTGCGCTGAAAGCGCTTAACCGTGACATCATCGC
CCAGCCGCGCAACGACGATCTTGCCGTTGGCGGCCTCGGCGGCGCGCTGTACCGCGAGCAGGTCGCCGTCGAGGATGCC
GGCATCGCGCATGCTCATGCCGCGCACTTTCAACAGGAAATCCGGCCGACTGGAAAACAGGGAAGGGTCGACCTGGTAT
TGCCGGTCGATGTGCTCGGCTGCCAGGATCGGGCTACCCGCCGCAACGCGGCCCACCAGCGGCAGCGTCAGCTGCATCA
GCCCCATCGACGGCAGCGAGAACTGGTGCGGCGATGCGCCGCCCTCCGCGCGCAGCCGGATACCGCGTGATGCGCCGGG
CGTCAGCTCGATCACGCCCTTGCGGGCGAGTGCCCGCAGGTGCTCCTCGGCCGCATTCGGCGACGAGAAGCCGAACTCC
```

By clicking the “FASTA” button on the Ribbon, any selected sequences will be printed in the FASTA format. You will be prompted to enter a header for your selected sequence(s). If multiple sequences are selected, the header you enter (plus a unique number to differentiate) will be used for all selected sequences. If a header is not entered, a default header of the form `> Seq #` will be generated for each sequence.

GENBANK

The **GenBank** format prints genetic sequences with 60 characters per line (6 groups of 10 characters, each group separated by a space). Each line is preceded by a number denoting the position of the first base in that line. Multiple GenBank sequences should be separated by two enters. By clicking the “GenBank” button on the Ribbon, any selected sequences will be printed in the GenBank format. An example of a sequence in GenBank format is show below:

```
  1 TTATGGAGCG GCTGGCCGGA TCAGGCCGAC CGCCAGTCCT TCCAACTGGA ACTCGTCGCG

 61 GTCGAGATCG ACGTGGATGG GTTCGAAATC AGGGTTCTCG GCAATCAGCT CGACCTGCCG

121 GCCTTTGCGC TGAAAGCGCT TAACCGTGAC ATCATCGCCC AGCCGCGCAA CGACGATCTT
```

REVERSE

This function **reverses** any selected sequence(s). It can be performed on both DNA/RNA and protein sequences. Below is a simple example:

```
>test sequence
TTATGGAGCGGCTGGCCGGATCAGGCCGACCGCCAG

>test sequence
GACCGCCAGCCGGACTAGGCCGGTCGGCGAGGTATT
```

COMPLEMENT

This function **complements** any selected sequence(s). It can only be performed on DNA/RNA sequences. If IUB characters are included (and allowed), they will not be affected by taking the complement. Below is a simple example:

```
>test sequence
TTATGGAGCGGCTGGCCGGATCAGGCCGACCGCCAG

>test sequence
AATACCTCGCCGACCGGCCTAGTCCGGCTGGCGGTC
```

REV. COMP.

This function takes the **reverse complement** of any selected sequence(s). It can only be performed on DNA/RNA sequences. If IUB characters are included (and allowed), they will not be affected by taking the reverse complement. Below is a simple example:

```
>test sequence
TTATGGAGCGGCTGGCCGGATCAGGCCGACCGCCAG

>test sequence
CTGGCGGTCGGCCTGATCCGGCCAGCCGCTCCATAA
```

TRANSLATION:

GENERAL OPTIONS (POP-UP)

Genetic Code

Standard

Vertebrate Mitochondrial

Yeast Mitochondrial

Mold, Protozoan and Coelenterate Mitochondrial

Invertebrate Mitochondrial

Ciliate, Dasycladacean and Hexamita Nuclear

Echinoderm and Flatworm Mitochondrial

Euplotid Nuclear

Bacterial, Archaeal and Plant Plasmid

BioWord allows the user to perform translations (codon -> amino acid) and reverse translations (amino acid -> codon) according to the nine genetic codes listed above.

FORWARD

This operation **translates** a DNA sequence into an amino acid sequence according to the selected *genetic code*. Only DNA sequences may be translated. If IUB characters are allowed, it is possible that multiple amino acids could satisfy a degenerate codon. BioWord will randomly select one of the possible amino acids. Remember if IUB characters are *not* allowed, they will be stripped from your sequence prior to the translation. This could interfere with the translation frame. Any "leftover" DNA bases (if the sequence length is not a multiple of 3 based on the *offset*) will be ignored.

Offset

0
1
2

The offset determines the reading frame for the translation.

```
0    ATC | GAG | GGC | CCG | A…
1    A | TCG | AGG | GCC | CGA…
2    AT | CGA | GGG | CCC | GA…
```

Below is a simple translation example, with the ***Standard Genetic Code*** and an offset of ***0***:

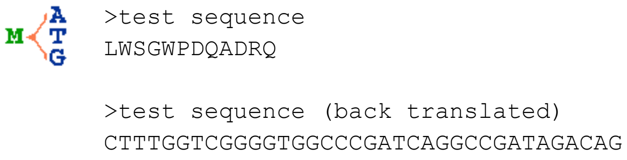
```
>test sequence
TTATGGAGCGGCTGGCCGGATCAGGCCGACCGCCAG

>test sequence (translated)
LWSGWPDQADRQ
```

REVERSE

This operation **reverse translates** an amino acid sequence into a DNA sequence according to the selected *genetic code* and method of reverse translation (described below). Only protein sequences can be reverse translated; an error message will be generated for any DNA/RNA sequences and they will be ignored.

Uniform

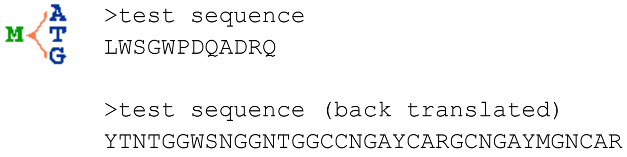
If the **uniform** option is selected, a random DNA codon will be selected from all of those that code for a given amino acid. Each codon has an equally likely chance of being selected.



```
>test sequence
LWSGWPDQADRQ

>test sequence (back translated)
CTTTGGTCGGGGTGGCCCGATCAGGCCGATAGACAG
```

IUB

If the **IUB** option is selected, each amino acid will be reverse translated into a single codon of degenerate IUB characters (according to the degenerate codons that code for that amino acid).



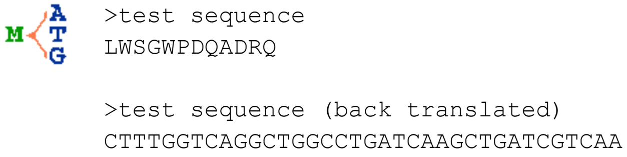
```
>test sequence
LWSGWPDQADRQ

>test sequence (back translated)
YTNTGGWSNGGNTGGCCNGAYCARGCNGAYMGNCAR
```

Best

If the **best** option is selected, the most frequently occurring DNA codon will be used for a given amino acid. The most frequently occurring codon will be determined by a **codon usage table**, which much be pasted into the prompt box. Codon usage tables in the proper format are available at http://www.kazusa.or.jp/codon/ using the option “a style like CodonFrequency output in GCG.”

Note that the prompt specifies NOT to include the column headings. This refers to the line highlighted in red below that is displayed on the above webpage.

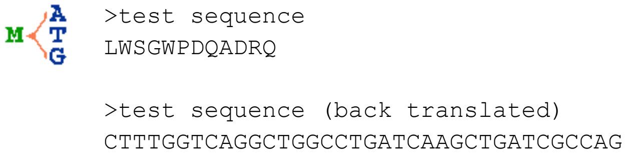
AmAcid	Codon	Number	/1000	Fraction	..
Gly	GGG	0.00	0.00	0.00	
Gly	GGA	1.00	2.06	0.00	
Gly	GGT	0.00	0.00	0.00	
Gly	GGC	10.00	20.58	0.00	



```
>test sequence
LWSGWPDQADRQ

>test sequence (back translated)
CTTTGGTCAGGCTGGCCTGATCAAGCTGATCGTCAA
```

Random Best

If the **random best** option is selected, the DNA codon for a given amino acid will be selected randomly via a weighted probability distribution function. Each codon is given a range of size proportional to its frequency as determined by a **codon usage table**. As described above, codon usage tables in the proper format are available at http://www.kazusa.or.jp/codon/using the option “a style like CodonFrequency output in GCG.”



```
>test sequence
LWSGWPDQADRQ

>test sequence (back translated)
CTTTGGTCAGGCTGGCCTGATCAAGCTGATCGCCAG
```

MAP

The **translation map** uses the selected *genetic code* to generate forward translations of a selected DNA sequence for every offset (0, 1, and 2). It handles IUB characters and leftover bases in the same manner as the *forward translation* operation. Each line is numbered with the position of the first base/amino acid in that line.

```
>test sequence
TTATGGAGCGGCTGGCCGGATCAGGCCGACCGCCAGATAGAGCCGGTACTCAACCCCCG

>test sequence
         1   M  E  R  L  A  G  S  G  R  P  P  D  R  A  G  T  Q  P  P
         1  Y  G  A  A  G  R  I  R  P  T  A  R  *  S  R  Y  S  T  P
         1 L  W  S  G  W  P  D  Q  A  D  R  Q  I  E  P  V  L  N  P
         1 TTATGGAGCGGCTGGCCGGATCAGGCCGACCGCCAGATAGAGCCGGTACTCAACCCCCG
```

DNA STATISTICS:

FREQUENCIES

This operation calculates the **frequencies** of A, U/T, C and G in a DNA sequence. IUB characters are not included in the base frequencies, but ARE included in the total length of the sequence.

```
>test sequence
TTATGGAGCGGCTGGCCGGATCAGGCCGACCGCCAGATAGAGC
```

Mononucleotide Frequencies	
Nucleotide	test seque
A	0.209
U/T	0.140
C	0.279
G	0.372

N-GRAM

The **N-gram** is a frequency representation of a DNA sequence. In essence, the DNA N-gram is generated by computing the histogram of the absolute frequencies of each N-nucleotide (di-, tri-, tetra-nucleotide). The N-grams will be listed in alphabetical order.

If IUB characters are allowed, any N-gram that contains an IUB character is not recorded in the counts. For example, for the sequence AATRRTC with an n-gram of 2, AA, AT, and TC would be recorded with counts of 1. TR, RR, and RT are ignored. If IUB characters are not allowed, the sequence would be read as AATTC and AA, AT, TT, and TC would be recorded with counts of 1.

Only display N-Grams in the sequence

If this option is selected, instead of listing every possible N-nucleotide, only those that are found in the sequence will be listed.

Include reverse complement

If this option is selected, the counts from the reverse complement of the selected sequence will be included in N-gram results.

Below is an example with N = 3, the "only display N-Grams in the sequence" option is checked and the "include reverse complement" option is not checked.

```
>test sequence
ATGATGGATGGCTATG
```

NGram (3) for Seq 'test sequence'	
NGram	Count
ATG	3
CTA	1
GAT	2
GCT	1
GGA	1
GGC	1
TGA	1
TGG	2

%GC GLOBAL

This operation calculates the **%GC** content of an entire selected DNA/RNA sequence. If IUB characters are allowed, they are included in the length of the sequence but not in the GC count.

```
>test sequence
TTATGGAGCGGCTGGCCGGATCAGGCCGACCGCCAGATAGAGC
```

%GC Content -> Global	
Position	test seque
0	0.651

%GC WINDOW

This operation computes **%GC** content by using a sliding window of a defined length. If IUB characters are allowed, they are included in the length of the sequence but not in the GC count.

Generate graph

If this option is selected, BioWord will generate a Microsoft Chart object graphing the %GC content as a function of window start position. If multiple sequences are selected, all will be graphed as different series on the same graph.

Length

Length is the size of the window over which %GC is computed.

Step-size

Step-size is how many bases the window skips when it is slid to the next position. A step-size of one will compute the %GC of a window starting at position 0, then at position 1, etc. A step-size of 5 will compute the %GC of a window starting at position 0, then at position 5, etc.

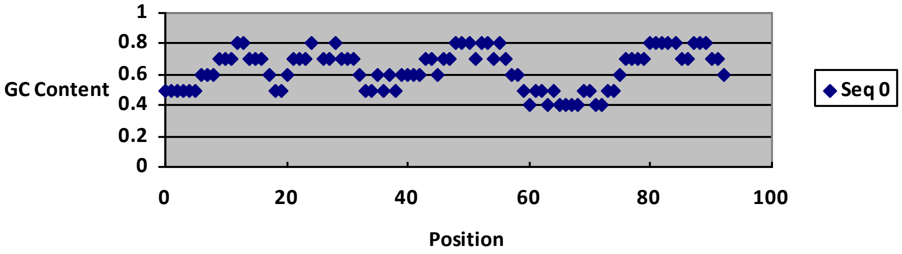
Below is an example with a window length of 10 and a step size of 1:

```
>test sequence
ATGATATGGCTATG
```

%GC Content -> Window Length: 10	Step Size: 1
Position	test seque
0	0.400
1	0.400
2	0.400
3	0.300
4	0.400

This example generates a graph for %GC on a longer sequence. This has a window length of 10 and a step size of 1:

```
 >Seq 0
 ATGCCAGGAATTCCCGGGGATCCTCCATCGGCGGAGTGTCGATGGACAGCGACCTCCGCCAGGAGATCATCCTCAAACCTTCCCGGA
CCCTCCCACCCCTAA
```



CODON USAGE TABLE

This operation generates a **codon usage table** according to a DNA sequence. It parses the sequence into codons, calculates frequency statistics, and uses the selected *genetic code* to determine corresponding amino acids. The codon usage table is printed in the “CodonFrequency output in GCG” format as found on http://www.kazusa.or.jp/codon/.

Offset

0

1

2

The offset determines the reading frame for identifying the codons.

Printing Options

Table – Results are printed as a Microsoft Word table object with cells/borders.

Whitespace – Results are formatted using tabs and returns rather than as an actual table

MOLECULAR WEIGHT

This operation calculates the **molecular weight** (in amu) of a DNA/RNA sequence. IUB characters, if allowed, are assigned an approximate weight according to the type of nucleic acid.

DNA

Double-stranded – Includes the weight of the sequence's complement strand

Include 5'-monophosphate – Adds 79 amu to each strand's weight

RNA

Include 5'-triphosphate – Adds 159 amu to the strand's weight

PROTEIN STATISTICS:

GENERAL

This operation calculates general **frequency** statistics for a protein sequence. It enumerates each individual amino acid as well as the sequence's composition of aliphatic, aliphatic hydroxyl, aromatic, sulfuric, basic, and acidic amino acids.

GRAVY

The **GRAVY** value for a protein sequence is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence.

ISOELECTRIC POINT

The **isoelectric point** of a protein sequence is the pH at which the protein does not have a net electric charge. The algorithm that BioWord uses can be found here (the bisection method): http://isoelectric.ovh.org/files/practise-isoelectric-point.html

MOLECULAR WEIGHT

This operation calculates the **molecular weight** (in amu) of a protein sequence. It is calculated by summing the average isotopic masses of amino acids in the protein and the average isotopic mass of one water molecule.

SUBSTRING SEARCHES:

ORFS

This operation finds **open reading frames** (ORFs) in a DNA sequence. An ORF begins with a start codon (BioWord accepts `ATG`, `CTG`, `GTG`, and `TTG` as start codons) and terminates with a stop codon (`TAG`, `TAA`, and `TGA`). BioWord will optionally search the reverse complement of a sequence for ORFs as well. If the *Replace Selection* option is selected, the optimal ORF will be highlighted in the sequence (if the ORF was in the original sequence, the highlight will be yellow; if it is in the reverse complement of the sequence, the highlight will be green.) If the *Below Selection* or *Save to Clipboard* option is selected, the frame, positions of the start and stop codon for the optimal ORF, its length in codons, and its sequence will be printed. If the *New Document* option is selected, all ORFs found in the selected sequence will be displayed.

Minimum Codon Length

The **minimum codon length** is the length threshold for an ORF. It includes the stop and start codons.

Include Reverse Complement

If this is selected, the reverse complement of the sequence will also be searched for ORFs. Any ORFs found in the reverse complement will be marked as such in the output.

Optimization

CAI + Length – If this option is selected, the ORFs are optimized by their CAI score times and their length in codons. CAI stands for **Codon Adaptation Index** and is a measure of codon bias according to a defined set of genes. This gene set is defined with a **codon usage table**. Codon usage tables in the proper format are available at http://www.kazusa.or.jp/codon/ using the option "a style like CodonFrequency output in GCG."

Length—If this option is selected, the ORFs are optimized by their length in codons.

The example below is optimized by length. The highlighted sequence was generated with the *Replace Selection* option and the printed results were generated with the *Below Selection* option.

```
>test sequence{Longest ORF: Start=2; Stop=103; Length=34;
Frame=1}
AATGCCAGGAATTCCCGGGGATCCTCCATCGGCGGAGTGTCGATGGACAGCGACCTCCGC
CAGGAGATCATCCTCAAACCTTCCCGGACCCTCCCACCCCTAAAAACAGATTTCTGTTTC

Longest ORF for 'test sequence'
Start Codon Pos: 2
Stop Codon Pos: 103
Length (in codons): 34
ATGCCAGGAATTCCCGGGGATCCTCCATCGGCGGAGTGTCGATGGACAGCGACCTC
CGCCAGGAGATCATCCTCAAACCTTCCCGGACCCTCCCACCCCTAA
```

SUBSTRING/GAPPED

BioWord has two options for string searching – **Substring Search** and **Gapped Search**. The **gapped search** allows the user to search for two strings separated by a gap of a defined length. The score of a gapped search match is the sum of the scores of the two strings.

For DNA/RNA sequences, the number of mismatches between the query sequence and selected sequence is calculated and used to judge the suitability of the match. If IUB characters are allowed, their mismatch penalties are calculated using weighted probabilities. For example, the IUB base R could be either A or G. If R is scored against A, it will have a mismatch penalty of 0.5. If R is scored against W (A or T), it will have a mismatch penalty of 1 – P(R, A) * P(W, A) = 0.75.

For protein sequences, there are two scoring methods. The desired scoring method is indicated in the *Advanced Options* menu. If the checkbox "use this instead of mismatches" is selected, the scoring matrix from the drop-down will be used to score (higher scores correspond to better matches). If the checkbox is not selected, a mismatch penalty system similar to DNA will be used.

There are two different ways to display results. If the *Replace Selection* option is selected, any matches will be highlighted in the sequence; the darker the highlight, the better the match. If any other output location is selected, the matches and their start position in the sequence will be printed in a table, ordered with better matches first.

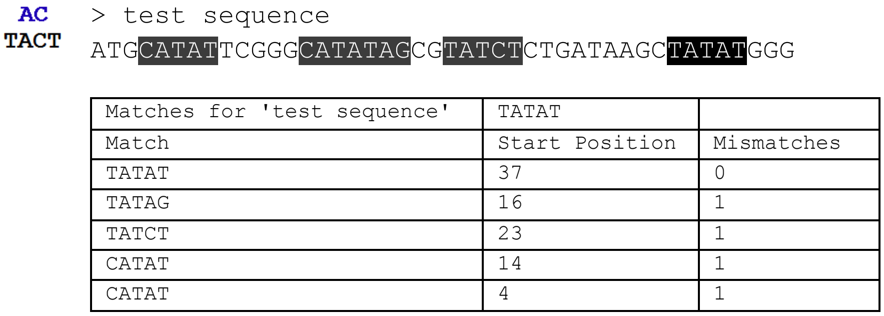
Mismatch Threshold

This option allows the user to define a **maximum threshold** for mismatches. If a substring of the sequence has a higher mismatch penalty than the threshold, it will not be included in the results.

Score Threshold

This option allows the user to define a **minimum threshold** for a score if a protein sequence is selected and the "use this instead of mismatches" checkbox is selected. If a substring has a lower score than the threshold, it will not be included in the results.

Below is an example of a **Substring Search** for TATAT with a mismatch threshold of 1. The highlighted result was generated when *Replace Selection* was selected and the tabulated result was generated when *Below Selection* was selected.



```
> test sequence
ATGCATATTCGGGCATATAGCGTATCTCTGATAAGCTATATGGG
```

Matches for 'test sequence'	TATAT	
Match	Start Position	Mismatches
TATAT	37	0
TATAG	16	1
TATCT	23	1
CATAT	14	1
CATAT	4	1

SITE SEARCH:

GENERAL OPTIONS

%GC Genome

This value is used to compute background DNA/RNA nucleotide frequencies for calculating $I_{sequence}$ and R_i sequence scores. It should be entered as an integer value. A 40% GC value implies a genome with 30% A, 30% T/U, 20% G and 20% C.

There are two options for background amino acid frequencies: a uniform distribution or the BLOSUM62 distribution. This is indicated in the *Advanced Options* menu.

Maximum Results

This value is used to limit the number of results printed.

IC Threshold

This option is used to define a score threshold, which is the value entered in the input box will be multiplied by the information content (IC) of the motif. Only results with an $I_{sequence}$ or R_i sequence score that is greater than the score threshold will be reported.

SEARCH

This operation uses a **motif** to **search** a selected sequence. It can be performed on both DNA/RNA and protein sequences. The sequences entered in the motif must be the same type and length and should be in FASTA format or separated by two enters. The scores are computed using R_i sequence or $I_{sequence}$, whichever is selected in the *Advanced Options* menu.

This operation generates a PSFM for the motif. IUB characters, if allowed, are factored into the PSFM based on weighted probabilities. For example, an R at a given position will increase the A count by 0.5 and the G count by 0.5.

If the *Replace Selection* option is selected, the matches will be highlighted in the sequence, with a darker highlight corresponding to a higher scoring match. If any other output location option is selected, the results, their positions in the sequence, and their scores will be listed in a table.

The example below uses the motif PGIP, PGCP, PGMP, PGIP, PGIP to search the selected sequence. The IC threshold is 0.3 X IC and the scoring function is R_i sequence. The highlighted result was generated with the *Replace Selection* option selected and the tabulated results were generated with the *Below Selection* option selected.

```
Ri      > test sequence
Iseq    LPGIPGDPPSAECRWTATSARRSSSNLPGMPSHP
```

Results for 'test sequence'		IC Threshold: 4.389
Result	Start Position	Ri Score
PGIP	2	16.072
PGMP	28	14.487

DYAD PATTERN

This operation uses a **dyad motif** to search a sequence. The two components of the dyad may be separated by a gap of defined length. The sequences entered in the motif must be the same type and length and should be in FASTA format or separated by two enters. The scores are computed using R_i sequence or $I_{sequence}$, whichever is selected in the *Advanced Options* menu. The score of a match is the sum of the scores of the two components.

This operation generates a PSFM for the motif. IUB characters, if allowed, are factored into the PSFM based on weighted probabilities. For example, an R at a given position will increase the A count by 0.5 and the G count by 0.5.

If the *Replace Selection* option is selected, the matches will be highlighted in the sequence, with a darker highlight corresponding to a higher scoring match. If any other output location option is selected, the results, their positions in the sequence, and their scores will be listed in a table.

Inverted Repeat/Direct Repeat Motif

This option determines the second half of the dyad. If the **inverted repeat** option is used, the second half will be scored using the reverse complement of the motif. If the **duplicate repeat** option is used, the second half will be scored using the same motif.

Below is an example using the motif TATG, TATC, TAAT, TTTC, TAGT, TATC, TATC with a gap ranging from 2-3 bases. The inverted repeat option is selected. The IC threshold is 0.3 X IC and the scoring function is R_i sequence. The highlighted result was generated with the *Replace Selection* option selected and the tabulated results were generated with the *Below Selection* option selected.

```
>test sequence
CTGTATCTTGATATTATTTCCGGTATCAAAGAAAGTAAG
```

Matches for 'test sequence'	(inverted repeat)	
Match	Start Position	Ri Score
TATCttGATA	4	12.9696516588714
TATCaaaGAAA	24	10.3846891581503

LOGO

This operation generates a **consensus sequence logo** for a sequence motif. A consensus logo is a visual representation of a motif. The base at each position is the base that occurs most frequently. The height of the bases is proportional to the position's conservation/information content. The motif's information content is also printed at the end of the logo. Each sequence must be the same length and of the same type. A consensus logo can be generated for both DNA/RNA sequences and protein sequences.

There are two ways to input sequences to create a logo:

1. Select a sequence motif in the document and then clicking the Logo button on the Ribbon.
2. With nothing selected, click the Logo button on the Ribbon, and input the sequence motif into the pop-up input box.

It is likely that the consensus logo will be printed at a very small font size. To scale the logo up or down, select it and click the *Bigger* and *Smaller* buttons in the Tools group on the Ribbon.

This operation generates a PSFM for the motif. IUB characters, if allowed, are factored into the PSFM based on weighted probabilities. For example, an R at a given position will increase the A count by 0.5 and the G count by 0.5.

Use IUB/Epsilon:

This is a DNA/RNA specific option. If two or more bases have frequencies within **epsilon** at a given position and this option is selected, the appropriate IUB character will be used in the consensus logo. If this option is not selected, the first encountered equiprobable base will be used in the logo.

Include PSFM

If this option is selected, a PSFM will be printed along with the logo. PSFMs can be printed in two formats:

Tabular PSFM—If this option is selected, a PSFM will be printed in a table format with columns for the alphabet letters and rows for the positions.

Jaspar Matrix—If this option is selected, a PSFM as used by the Jaspar database will be printed. The Jaspar matrix format is as follows, with each number representing the amount of times that base was found in that position:

```
A  [13 13  3  1 54  1  1  1  0  3  2  5 ]

C  [13 39  5 53  0  1 50  1  0 37  0 17 ]

G  [17  2 37  0  0 52  3  0 53  8 37 12 ]

T  [11  0  9  0  0  0  0 52  1  6 15 20 ]
```

Scale (bits)

This value defines the size (in bits) of the vertical reference line that precedes consensus logos.

The example below has *Use IUB* and *Include PSFM* (tabular) selected and has a scale of 2 bits.

```
Selected Motif:
TTTCTGGT    CCTATTGT
TCAATTGT    TCCATTGT
CTCATTGA    CCAATTGT
CCCATTGT    TCCATTGT
```

PSFM	A	C	G	T
1	0.000	0.500	0.000	0.500
2	0.000	0.750	0.000	0.250
3	0.250	0.500	0.000	0.250
4	0.875	0.125	0.000	0.000
5	0.000	0.000	0.000	1.000
6	0.000	0.000	0.125	0.875
7	0.000	0.000	1.000	0.000
8	0.125	0.000	0.000	0.875

(2 bits) | YCcATTGT [IC: 11.058 bits]

MOTIF DISCOVERY:

GIBBS/GREEDY SAMPLING

These operations find **optimal motifs of defined length** within a collection of sequences. The **Gibbs Sampling** algorithm works by randomly selecting a substring of defined length from each sequence in the collection, and adding them to a temporary motif. One sequence is excluded. The motif and excluded sequence are scored against each other using R_i sequence or $I_{sequence}$ via a sliding window. The scores are used to generate a cumulative distribution function. Once all valid positions in the excluded sequence have been tested, a random number is generated to determine which window of the excluded sequence will be added to the temporary motif for the next iteration. In **Gibbs Sampling**, this window is randomly selected using the cumulative distribution function. In **Greedy Sampling**, the window is the one found to have the highest score against the motif. The selected window is added to a motif and another sequence is excluded. This continues until the algorithm has repeated the specified number of times.

The entire process is repeated ten times, and the motif with the highest information content will be kept as a result. The user can opt to generate multiple motifs; in this case, this process will be repeated for each result. BioWord uses masking to improve the chances of finding alternate/weaker motifs. Results will be ordered by information content. Please note that this operation may take time to complete.

This operation generates a PSFM for the sample motifs. IUB characters, if allowed, are factored into the PSFM based on weighted probabilities. For example, an `R` at a given position will increase the `A` count by 0.5 and the `G` count by 0.5.

If the *Replace Selection* option is selected, the optimal motif will be highlighted in each sequence in the alignment. If any other output location option is selected, the sequences that make up the optimal motif and their positions in each sequence will be printed in a table.

%GC of Genome

This value is used to compute background DNA/RNA nucleotide frequencies for calculating $I_{sequence}$ and R_i sequence scores. It should be entered as an integer value. A 40% GC value implies a genome with 30% `A`, 30% `T`, 20% `G` and 20% `C`.

There are two options for background amino acid frequencies: a uniform distribution or the BLOSUM62 distribution. This is indicated in the *Advanced Options* menu.

Number of Iterations

This value is the number of times to run the Gibbs/Greedy Sampling algorithm in the hopes of discovering the best motif.

Window Length

This value is the length of the motif to search for in the collection. It must be less than or equal to the length of the shortest sequence in the selected collection.

Maximum Results

This value is used to limit the number of results printed.

Masking Parameter

To ensure that all results are unique, BioWord masks the positions of each discovered motif in each of the sequences. The masking parameter allows the user to additionally mask positions surrounding the ones that define the start of the motif. For example, consider the following parameters: window-length = 8, masking parameter=0.25. If the motif was found at position 20, BioWord will also mask within 8 * 0.25 = 2 positions of 20. In this case, positions 18, 19, 20, 21 and 22 will be masked.

This **Greedy Sampling** example has a window length of 10 and was run 100 times. The highlighted sequences were generated with the *Replace Selection* option (**Greedy Sampling** highlights in green and **Gibbs Sampling** highlights in yellow) and the tabulated results were from the *Below Selection* option.

```
MSGERCFGRICDKRSSPNAAEEHLKALARKGVIEIVSGASRGIRLLMEEEPSEGLPLIGRVAAGEPLLA

FKGIRAAQYHLEALEHAGAIRRVPGQARGIRLAGQGAQTRTAPVSEVARDDVLRLPVLGRVAAGLPIGDSMRDEGIF

MLKLTPRQAEILAFIHIEQSCNINPAFFHPQADYLLRVHGMSMKDVGIFDGDLLAVHTC

MKALTARQQEVFDLNAAEEHLKALARKGVLEIVSGASRGIRLLQEEEDGLPLVGRVAAGEPLLAQQHIEGHYQVDPS

MKPLTAREEGLPLIGQVAAGEPILAEQHVEGTYKVDPNMFKPQADFLLKVYGQSMKDIGILDGDLLAVHST

MTLSRNNNAKRGLQLSQRKVVAPAATSPAFELPLVGIVAAGRPVEAFQLSDMDGDFVAVHPQ

MLTRIKALEERGFIRRLPNRARALEVIRLPENRTDSNQQEKKVRENFSLPKAHNDVVELPLHGRIAAGLPIGD
```

Greedy Sampling Motif:	IC = 32.8026088258276	
Sequence	Aligned	At Position
Seq 0	RVAAGEPLLA	60
Seq 1	RVAAGLPIGD	60
Seq 2	EILAFIHIEQ	10
Seq 3	RVAAGEPLLA	55
Seq 4	QVAAGEPILA	16
Seq 5	IVAAGRPVEA	37
Seq 6	RIAAGLPIGD	64

DYAD MOTIF

This operation searches a sequence for all **dyad motifs** of a defined length with a defined spacer. The score of a match is the sum of the scores (or mismatches) of the two components. The operation can be performed on DNA/RNA and protein sequences.

For DNA/RNA sequences, the number of mismatches between a perfect dyad and a potential dyad from the sequence is calculated and used to judge the suitability of the match. If IUB characters are allowed, their mismatch penalties are calculated using weighted probabilities. For example, the

IUB base R could be either A or G. If R is scored against A, it will have a mismatch penalty of 0.5. If R is scored against W (A or T), it will have a mismatch penalty of 1 – P(R, A) * P(W, A) = 0.75.

For protein sequences, there are two scoring methods. The desired scoring method is indicated in the *Advanced Options* menu. If the checkbox "use this instead of mismatches" is selected, the scoring matrix from the drop-down will be used to score a potential dyad (higher scores correspond to better matches). If the checkbox is not selected, a mismatch penalty system similar to DNA will be used.

Dyad Length

This value determines the length of one component of the dyad to look for in the selected sequence. BioWord allows the user to specify a range of dyad lengths.

Spacer Length

This value determines the length of the spacer to allow between dyad components. BioWord allows the user to specify a range of lengths for the spacer.

Mismatch Threshold

This option allows the user to define a **maximum threshold** for mismatches. If a dyad motif found in the sequence has a higher mismatch penalty (scored against a perfect dyad) than the threshold, it will not be included in the results.

Score Threshold

This option allows the user to define a **minimum threshold** for a score if a protein sequence is selected and the "use this instead of mismatches" checkbox is selected. If a dyad motif has a lower score than the threshold (scored against a perfect dyad), it will not be included in the results.

Maximum Results

This value is used to limit the number of results printed.

Inverted Repeat/Direct Repeat

This option determines the desired sequence of the second half of the dyad. With the **inverted repeat** option, the second half of the dyad will be judged against reverse complement of the first half. With the **duplicate repeat** option, the second half of the dyad will be judged against the first half.

Below is an example searching for dyads (direct repeat) with length 4 ± 1, separated by a spacer with length 3 ± 1. The mismatch threshold is 2 and maximum results are limited to 6.

```
>test sequence
TACGAGATCATTACATTGCATATTTTAGGGATCTATGCTCATTGGGCTATATA
```

Dyads for 'test sequence'	(direct repeat)	

Dyad	Start Position	Mismatches
CATtaCAT	9	0
CATtgCAT	14	0
ATTgcatATT	15	0
ATTacATT	10	0
ATGctcaTTG	35	1
TAGGgatcTATG	26	1

ALIGNMENT:

BioWord can align DNA/RNA and protein sequences using a global or local algorithm (see respective sections for more details). For printing output, if the *FASTA* output format is selected, the sequences will be printed separately in FASTA format with gaps corresponding to the calculated alignment. However, if *Raw* or *GenBank* output formats are selected, BioWord will produce a more BLAST -esque alignment output. For protein sequences, BioWord marks conservation groups with a "+" symbol.

See http://www.cgl.ucsf.edu/chimera/docs/ContributedSoftware/multalignviewer/clustgroups.html for a description of amino acid conservation groups.

GENERAL OPTIONS

Gap Opening Penalty (GOP)

This is the value associated with opening a gap in an alignment.

Gap Extension Penalty (GEP)

This is the value associated with extending a gap in an alignment.

Maximum Results

This value is used to limit the number of results printed.

DNA SPECIFIC OPTIONS

Match/Mismatch

This option allows the user to define the score assigned to a (string) match and of a (string) mismatch. That is to say that A with A would be a match, and A with any character other than A would be a mismatch.

Matrix Scoring

This option allows the user to be more specific with scoring. The values in the matrix correspond to the score between two characters. If an IUB character is allowed and encountered, its score is 0 regardless of the other character.

GLOBAL (NEEDLEMAN-WUNSCH)

BioWord's **pair-wise global alignment** is based on the Needleman-Wunsch algorithm. This method attempts to align the entirety of both sequences. Global alignments are useful when both sequences are similar on the whole. This function can be performed on both DNA/RNA and protein sequences. Protein sequences are aligned using the scoring matrix selected in the *Advanced Options* menu.

The example below uses the following values: GOP = -3, GEP = -2, match = 2, and mismatch = -1. The output format is *Raw*.

```
>test 1
TGGTAGATTCTACCGAAACCCCAAATATATAGGTAGGGGGACGTTCGCGGATGGTATAGATGT

>test 2
GCCATCGGCCGGGTGAATTGCGAGTAATAAACCCCAAATATACAGCGGTACGGGGGGTATATATATAT

Global
Match=2; Mismatch=-1
GOP=-3; GEP=-2
       1 ----T------GGTAGATT-CTACCGA--AACCCCAAATATATAGGTAGGGGGACGTTCG
             |      |||  ||| | |   |  ||||||||||||| ||     || |||   |
       1 GCCATCGGCCGGGTGAATTGCGAGTAATAAACCCCAAATATACAGC----GGTACG---G

      61 CGGATGGTATAGATGT
          || |  |||| || |
      61 GGGGTA-TATATATAT
```

LOCAL (SMITH-WATERMAN)

BioWord's **pair-wise local alignment** is based on the Smith-Waterman algorithm. This method attempts to find a local optimal alignment. Local alignments are useful when two sequences are overall different, but are suspected to contain regions of similarity. This function can be performed on both DNA/RNA and protein sequences. Protein sequences are aligned using the scoring matrix selected in the *Advanced Options* menu. The position where the optimal alignment begins is added to each sequence's header.

The example below uses the following values: GOP = -3, GEP = -2, match = 2, and mismatch = -1. The output format is *FASTA.*

```
>test 1
TGGTAGATTCTACCGAAACCCCAAATATATAGGTAGGGGGACGTTCGCGG

>test 2
GCCATCGGCCGGGTGAATTGCGAGTAATAAACCCCAAATATACAGCGGTACGGG


>test 1 {Local; Match=2; Mismatch=-1; GOP=-3; GEP=-2; Pos=2}
GGTAGATT-CTACCGA--AACCCCAAATATATAG--GTAGGGG

>test 2 {Local; Match=2; Mismatch=-1; GOP=-3; GEP=-2; Pos=12}
GGTGAATTGCGAGTAATAAACCCCAAATATACAGCGGTACGGG
```