TAMP Finder

Application

Overview

January 2019

CRC Systems Inc
333 Plains Rd W
Burlington, ON L7T 1G1
905-639-8725
www.crcsystems.com
tamp@crcsystems.com

1. Introduction

Tail-anchors are hydrophobic sequences located at the carboxyl-terminus of a variety of functionally unrelated proteins. These sequences serve to target the protein to the appropriate subcellular membrane as well as to anchor it into the lipid bilayer.

In order to be identified as having a tail-anchor, a protein must have a region of hydrophobicity near the C terminus, yet it cannot have a significant region of hydrophobicity near the N terminus, or in the center of the protein.

2. Application Overview

The TAMP application was designed to computationally evaluate protein sequences and determine if they meet the definition of a tail anchor protein, given a set of user entered criteria.

The application is written primarily in Perl, and it stores the data in an SQL database. The user interface is web-based and written in PHP.

2.1 Defined Regions

The protein sequence is divided into three regions: N terminal, Center and C terminal regions. The length of each region is dependent on the length of the protein, and is user defined. See **Table 4** for the default values.

Four evaluations using a set of user defined criteria are then performed on each of the three regions.

2.2 Application Dependencies

Table 1: List of user definable parameters used by the application for each of the four evaluations.

Parameter	Definition
Scan length	The user defined number of amino acids that the application will scan in each region. The scan length is also dependent on the length of the protein as set out in **Table 2**. See **Table 4** for the default settings.
Window size	The number of amino acids that are averaged at one time. The application defines the head and the tail as the first and last amino acid of the window, respectively. The window size is defined by the user for each region, see **Table 4** for the default settings.
Window weight	The multiplication factor given to each amino acid position in the window. The number of weightings available is equal to the window size. These values are user defined, see **Table 4** for the default settings.
Scale	The hydrophobicity scale that the application uses in each region to give a numerical value to each amino acid for use in the averaging of the window. The user can choose from a number of pre-defined scales, or create a scale of their own. The pre-defined scales include: Kyte-Doolittle, Hopp-Woods, Engleman-Steitz, White, Janin, Chothia, and Eisenberg-Weiss. For non-standard amino acids selenocysteine (U) and glutamine or glutamic acid (Z), and for unknown (X), the median value of the scale is used.
Contiguous hits	The number of consecutive averaged window values required to determine a region of hydrophobicity. This value is user defined, see **Table 4** for the default settings.
Thresholds	The value that each contiguous hit must reach in order to determine whether the hit is hydrophobic **OR** hydrophilic enough for the given test. There are two thresholds for each test, and both are user defined, see **Table 4** for the default settings. The two thresholds must be satisfied in the same stretch of contiguous region.
Comparator	Greater than or equal to **OR** less than or equal to, comparison of the value of the hit against the threshold. This value is user defined, see **Table 4** for the default settings.

2.3 Overview of Evaluation

C test searches the C terminal region of the protein for a region of hydrophobicity. A region of hydrophobicity is required in the C test in order for it to be a Tail Anchor protein.

N test searches the N terminal region of the protein for a region of hydrophobicity. A region of hydrophobicity in the N test indicates the presence of a signal peptide, and therefore is not a Tail Anchor protein.

Center test searches the region in between the N terminal region and C terminal region for a region of hydrophobicity. A region of hydrophobicity in the Center test indicates that the protein is not a Tail Anchor protein.

C2 Test is a repeat of the **C test**, using an alternate scale and parameters. The purpose of repeating the test using a different scale is to determine the core and borders of the hydrophobic region in the C terminal region.

2.4 Application Results

A Tail Anchor protein is positively identified if it has a region of hydrophobicity in both the C test and the C2 test, and no region of hydrophobicity in the N test and Center test. The *pass method* given to sequences in this category is the classification name of the protein, which is dependent on sequence length: CP, CP1, AP1, AP2. See **Table 2** for classification values.

Example: CP indicates that the test found two overlapping regions of hydrophobicity using both the C test and C2 test, no region of hydrophobicity in the N test or Center test, and the protein sequence length is greater than or equal to 119 amino acids.

If a protein is found to have a region of hydrophobicity in either the C test or C2 test, and no regions of hydrophobicity in N test or Center test, then it is labeled as a one-pass protein (OP). The 2 letter abbreviation of the scale that found the region of hydrophobicity is appended to the end of the label for annotation, along with the classification name.

Example: OPHWCP indicates the test only found a region of hydrophobicity using the Hopp Woods scale, and the sequence was evaluated using the parameters for CP proteins (and therefore is greater than or equal to 119 amino acids long).

2.5 Borders Calculation

Border calculations are performed on all sequences that are evaluated as having a Tail Anchor region. These sequences are divided into two distinct groups. One group is the One Pass proteins (OP) and the other group is the Complete (CP, CP1) or Auto Pass proteins (AP1, AP2). The One Pass protein border calculations are straight forward. The NTS (N-Terminal Sequence) is the 15 amino acids after the N border where the N border is only the first amino acid outside the TMS (Transmembrane Sequence) on the N terminus side of the sequence. The C border is the first amino acid outside the TMS on the C terminus side of the sequence and the CTS (C-Terminal Sequence) is the rest of the sequence after the C border on the C terminus side. The TMS is deemed to be the first sequence of amino acids from the C terminus side that fulfills the contiguous hits and contains at least one hit that is greater or less than the defined threshold.

For Complete and Auto Pass protein border calculations, there are two scenarios that need to be considered. The first scenario is when the two hydrophobic regions calculated using the two separate scales overlap. The second scenario is when the pass method shows that the sequence is a complete pass but the hydrophobic regions from the two scales do not overlap. In the first scenario where the two hydrophobic regions overlap, the N border is where the two scales do not overlap on the N terminus side of the TMS and the C border is where the two scales do not overlap on the C terminus side of the TMS.

E**xample 1**:

```
<-N Term                                C Term
MGAIGAVRCSSSRSLGPSGNVPPPPSAPAPGKNEWGTDAPRLTVA
FFFFFFFFFFFFFFFFFF22222111112222222222FFFFFFF
FFFFFFFFFFFFFFFFF222211111111111122FFFFFFFFFF

NTS = AIGAVRCSSSRSLGP
N border = SG
TMS = NVPPPPSAPAPGKNEW
C border = GTD
CTS = APRLTVA
```

The second scenario is when the two scales do not have any overlap. In this case the first contiguous hits of either scale from the C terminus is deemed the TMS core.

Example 2:

```
<-N Term                                                            C Term
MGAIGAVRCSSSRSLGPSGNVPPPPSAPAPGKNEWGTDAPRLTVAVPPPPSAPAPGKNEWGTDAPRLP
FFFF2222111111222222FFFFFFFFFFFFFFFFFFFFFF22222111111122211FFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFF22222111111222222FFFFFFFFFFFFFFFFFFFFFFFFFFFF

NTS = APAPGKNEWGTDAPR
N border = L
TMS = TVAVPPPPSAPAPGKNE
C border = W
CTS = GTDAPRLP
```

Other special cases include where the TMS runs all the way to the end of the C terminus, which means that there is no C border or CTS. If the TMS runs to the second last amino acid from the C terminus, then there is a C border, which is the last amino acid but no CTS.

If the TMS runs into the Center region, the NTS and N borders are still calculated the same way. Additionally, the TMS core can continue into the Center region as long as the calculated window average satisfies the defined C terminal thresholds.

If a protein is found to have no region of hydrophobicity in the N test and Center test, but found to have a region of hydrophobicity in both the C test and C2 test, without either region overlapping, then it is labeled as a double one pass protein, with the 2 letter abbreviation of the scale appended to the end for annotation, along with the classification name.

Example: OPHWAP1OPKDAP1 indicates the test first found a region of hydrophobicity using the Hopp Woods scale, then found another region of hydrophobicity using the Kyte-Doolittle scale, and that the sequence was evaluated using the AP1 set of parameters, and the sequence is between 79 and 103 amino acids in length.

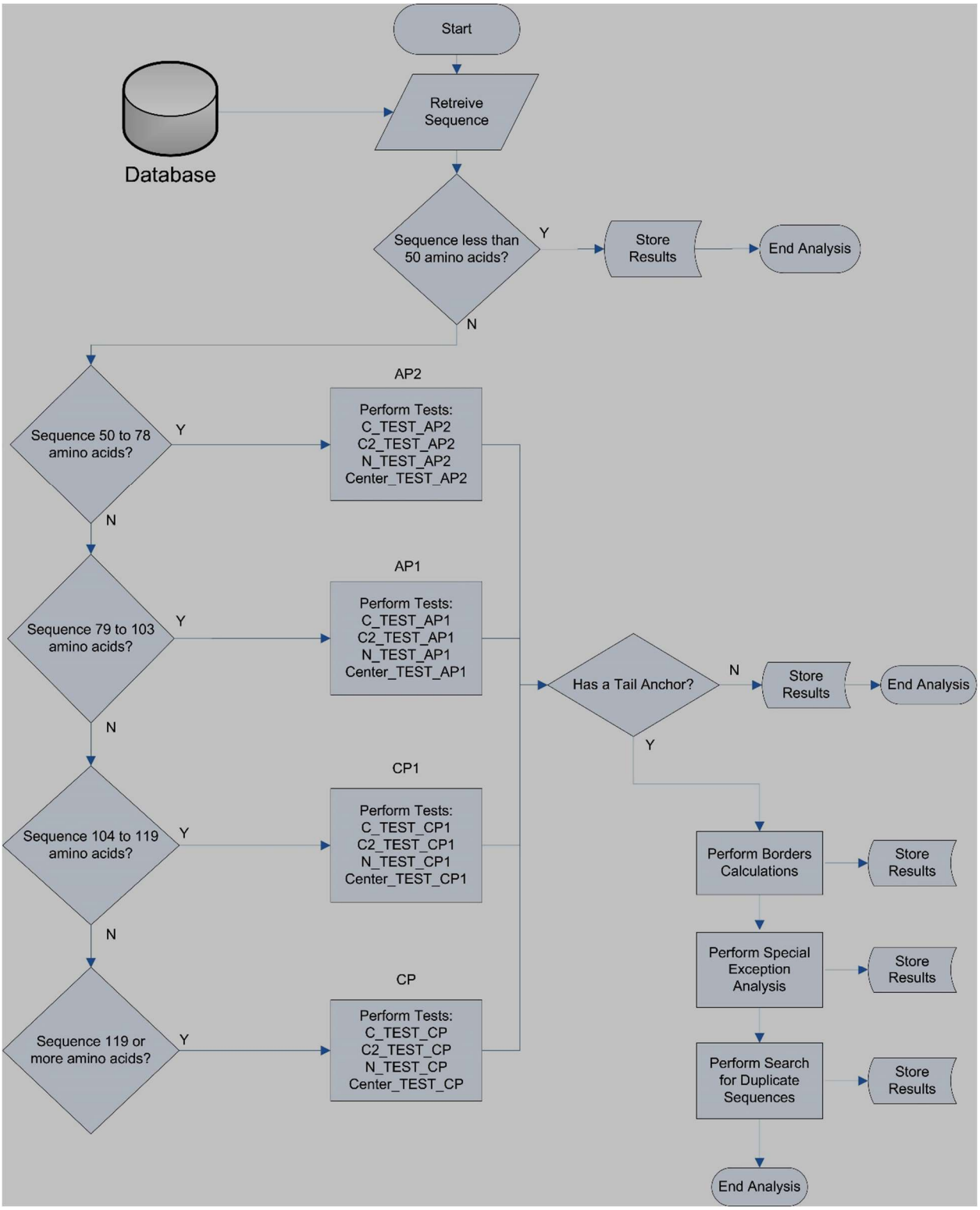
If a protein is found to have a region of hydrophobicity in both the C test and C2 test and an additional region(s) of hydrophobicity, the number of additional hydrophobic regions is annotated along with the scale that was used to determine the region. This information is recorded in *border_info*. The first region of hydrophobicity encountered, when scanning from the C to N terminus, is used for the border calculations.

2.6 Duplicates Search

A duplicate protein has a sequence or region of a sequence that is partially identical to another protein. After the program determines which sequences contain Tail Anchor regions, the program then searches the database for duplicate proteins. Duplicate proteins in this application are defined as having the last 10 amino acids 100% identical, and amino acids 11 to the start of the C terminal region from the carboxy terminus as 90% identical. When comparing the sequences, the shorter scan length of the two proteins being compared is used for the calculation of similarity. The duplicate with the longest sequence length is the "parent", and is included in results. The shorter duplicates are flagged in the "duplicateOf" field with the parent's accession number and are not displayed in the results (unless otherwise selected in the user interface).

3. Program Details:

Each protein sequence is evaluated in the following method:



3.1 Classifications

The protein is divided into four classifications: CP, CP1, AP1, AP2, depending on the length of the sequence. See **Table 2** for the default values.

Any sequence less than 50 amino acids, is not tested, and determined to not have a Tail anchor region.

The values from **Table 2** were computationally determined to be optimal by experimentation. Each of these values can be set by the user prior to running the program, with the default values specified in **Table 4**.

Table 2: Default values of the N terminal, C terminal and Center scan lengths, are dependent on the sequence length. These values can be defined by the user.

Classification	**Sequence Length >=**	**N Terminal Scan Length (amino acids)**	**C Terminal Scan Length (amino acids)**	**Center Length (amino acids)**
CP	119	40	65	sequence length – (40 + 65)
CP1	104	40	50	sequence length – (40 + 50)
AP1	79	20	45	sequence length – (20 + 45)
AP2	50	20	30	sequence length – (20 + 30)

3.2 Regions

The sequence is divided into 3 regions (N terminal, Center, C terminal) depending on the protein length, as defined in **Table 2**.

3.3 Parameters

The program parameters for each region are selected by the user. The parameters as detailed in **Table 1** include: scan length, window size, window weight, scale, contiguous hits, threshold 1, threshold 2, and comparator, for each CP, CP1, AP1, AP2 sequence, in each of the C test, N test, Center test, and C2 test.

3.4 Order of Tests

3.4.1 The **C test** searches the C terminal region of the protein for a region of hydrophobicity. The application uses the selected hydrophobicity scale and converts each amino acid in the sequence into the scale's corresponding number. The numbers are averaged across the window, and that number is stored. The starting position for the head of the C test is calculated as: sequence length – (scan length - 1) – (window size - 1). The tail is calculated as: head + window size -1. See **Table 3** for summary of the scan window formula.

The window moves across the sequence one position at a time towards the carboxy terminus, and stores the value of the average window. The scan runs while the tail is less than the sequence length. If enough contiguous values meet the thresholds, then the region is determined to contain a region of hydrophobicity. A region of hydrophobicity is required in the C test in order to be identified as a Tail Anchor protein.

3.4.2 The **C2 test** is a repeat of the C test, using an alternate scale and parameters. A second evaluation of the C terminal region is performed in order to more accurately calculate the hydrophobic region core and borders.

3.4.3 The **N test** searches the N terminal region of the protein for a region of hydrophobicity. The application uses the selected hydrophobicity scale and converts each amino acid in the sequence into the scale's corresponding number. The numbers are averaged across the window, and that number is stored. The window scan starts where the head of the window is on the first amino acid of the sequence. The tail is (head + window size -1). The window moves across the sequence one position at a time, and records the value of the average window. The scan stops when the head of the window is on the last amino acid of the N terminal region. If enough contiguous values meet the thresholds, then the region is determined to contain a region of hydrophobicity. A region of hydrophobicity in the N test indicates the presence of a signal peptide, and therefore indicates that the sequence is not a Tail Anchor protein. See **Table 3** for summary of scan window formula.

3.4.4 The **Center test** searches the region in between the N terminal region and C terminal region for a region of hydrophobicity. The application uses the selected hydrophobicity scale and converts each amino acid in the sequence into the scale's corresponding number. The numbers are averaged across a window, and that number is stored. The window scan head starts on the first amino acid after the N terminal region, and the tail is (head + window size -1). The window moves across the sequence one position at a time, and records the value of the average window. The scan stops where the last position of the window is one amino acid before the start of the C terminal region (while head < =(sequence length – C scan length – window)). If the defined number of contiguous hits meets the thresholds, then the region is determined to contain a region of hydrophobicity. A region of hydrophobicity in the Center test indicates that the protein is not a Tail Anchor protein. See **Table 3** for summary of scan window formula.

Table 3: Positions and scan of the window in each of the tests

Test	Head	Tail	Scan Until
C Test	sequence length – (scan length - 1) – (window size - 1)	head + window size -1	while tail <= sequence length
N Test	first amino acid in sequence	head + window size -1	while head <= N scan length
Center Test	N scan length	head + window size -1	while head <= (sequence length – C scan length – window size)
C2 Test	sequence length – (scan length - 1) – (window size - 1)	head + window size -1	while tail <= sequence length

4. User-Defined Default Program Parameters

The following table shows the parameters that were computationally determined to be optimal. CP proteins are greater than or equal to 119 amino acids, CP1 proteins are between 104 and 118 amino acids, AP1 proteins are between 79 and 103 amino acids, and AP2 proteins are between 50 and 78 amino acids. See **Table 2** for details on the classification and default lengths.

Example: If the protein is 125 amino acids long, the parameters in the table labeled as C_TEST_CP, N_TEST_CP, CENTER_TEST_CP, C2_TEST_CP. Likewise, proteins that are 55 amino acids long will use the parameters for the test name ending in C_TEST_AP2, N_TEST_AP2, CENTER_TEST_AP2, C2_TEST_AP2.

Table 4: Default application parameters. All comparator values are interpreted by the application as <= or >=.

Test Name	Window Size	Scan Length	Scale Name	Weight	Threshold A	Threshold B	Comparator	Contig Hits
C_TEST_CP	9	65	Hopp-Woods	1,1,1,1,1,1,1,1,1	-0.96	-0.6	<	12
C_TEST_CP1	9	50	Hopp-Woods	1,1,1,1,1,1,1,1,1	-0.96	-0.6	<	12
C_TEST_AP1	9	45	Hopp-Woods	1,1,1,1,1,1,1,1,1	-0.96	-0.6	<	12
C_TEST_AP2	9	30	Hopp-Woods	1,1,1,1,1,1,1,1,1	-0.96	-0.6	<	12
N_TEST_CP	9	40	White	1,1,1,1,1,1,1,1,1	0.3	0	>	6
N_TEST_CP1	9	40	White	1,1,1,1,1,1,1,1,1	0.3	0	>	6
N_TEST_AP1	9	20	White	1,1,1,1,1,1,1,1,1	0.3	0	>	6
N_TEST_AP2	9	20	White	1,1,1,1,1,1,1,1,1	0.3	0	>	6
CENTER_TEST_CP	8	0	Engleman-Steitz	1,1,1,1,1,1,1,1	-2.1	-1.7	<	7
CENTER_TEST_CP1	8	0	Engleman-Steitz	1,1,1,1,1,1,1,1	-2.1	-1.7	<	7
CENTER_TEST_AP1	8	0	Engleman-Steitz	1,1,1,1,1,1,1,1	-2.1	-1.7	<	7
CENTER_TEST_AP2	8	0	Engleman-Steitz	1,1,1,1,1,1,1,1	-2.1	-1.7	<	7
C2_TEST_CP	7	65	Kyte-Doolittle	1,1,1,1,1,1,1	2	0.8	>	12
C2_TEST_CP1	7	50	Kyte-Doolittle	1,1,1,1,1,1,1	2	0.8	>	12
C2_TEST_AP1	7	45	Kyte-Doolittle	1,1,1,1,1,1,1	2	0.8	>	12
C2_TEST_AP2	7	30	Kyte-Doolittle	1,1,1,1,1,1,1	2	0.8	>	12