

## Introduction

AminoTrack™ is a web based program designed to increase the efficiency which sequence data is recorded for further analysis with statistical programs such as SPSS. The main purpose of AminoTrack™ is to streamline the process and reduce human error in the identification of mutations present in multiple sequences as compared to a reference or wild-type sequence. A set of aligned protein sequences is input in the web submission form and comma delimited output files are generated as a zip file for loading into a spreadsheet format. These files can be opened or imported into any spreadsheet program that recognizes comma delimited files (Microsoft Excel, SPSS, etc.). The sequences are analyzed for mutations in amino acids, charge changes, potential N-linked glycosylation sites, insertions, and deletions. The data is recorded in a spreadsheet in a columnar binary format of “0” and “1” with each amino acid in its own column with a position number at the top, with “0” indicating no mutation, and “1” indicating a mutation. The program also generates similar outputs for charge changes and lists the amino acid N-linked glycosylation motifs for each site present. The position numbers of each residue is calculated by the program based on the user’s input in the sequence submission screen. Currently this program is being used to identify mutations in the protein sequences of the outer Envelope protein during HIV and SIV infection.

## Input File – Protein sequences

1. Align your sequences. The program was developed for use with Clustal “.gde” or FASTA text only files. Edit the alignment to remove extra gaps and misalignments. The file should be in FASTA format; sequence names should be denoted by either “>”, “#”, or “%”.
2. Copy and paste the aligned sequences in the AminoTrack™ sequence input window.
3. Enter starting amino acid position number and name for the data run in the appropriate boxes. The starting position number is the position number of the first amino acid in your reference sequence. All other residues will be numbered in relation to this starting position.
4. Click generate data
5. Save and extract output files from the zip file.

## Output Files

All files can be opened in Excel and SPSS, and any other program that recognizes comma delimited files in a spreadsheet format. There are 5 types of output files as listed below. All types may have multiple files if the sequences are long since the data needs to be split-up due to limitations in the number of columns per file in Excel. For example a file “AASeqChanges 1-398” indicates that only information for amino acid residues 1-398 is included in this file. If the sequences are longer, then separate file(s) will contain the rest of the information.

1. AASeqChanges#-#:

These files list the entire sequence and positions at which the sequences differ from a reference sequence. For each position, no change from the reference is represented as a “-“, while mutations are indicated by the letter of the mutant residue or a “.” in the case of a deletion.

## 2. Charge:

### Charges#\_#:

These files list amino acid positions that result in a charge change. The charge for each charged residue is listed as -1 (D, E) or 1 (R, K). The values are listed for each sequence including the reference sequence. The total charge change in relation to the reference is not automatically calculated but can be done so manually in Excel for each sequence. For example if there is a K (reference sequence) to E (mutant sequence) change, the values listed will be +1 for the reference and -1 for the mutant.

### ChargeChanges#\_#:

This is similar to the charge outputs above except that the individual charge values for each residue are not given. At positions that have a mutation affecting charge, the output lists the resulting change in charge from the reference sequence. For example, if there is a K (reference sequence) to E (mutant sequence) mutation at position 100, the charge changes from 1 (K) to -1 (E). The value listed at position 100 will be -2, since there is a net change of -2. Similarly for a E (-1) to K (1) change, the value listed will be +2.

Also if D or E (-1) changes to an uncharged residue such as N, the value listed will be +1. This is because the charge changed from -1 to 0 for a net change of +1. Similarly if R or K (+1) changes to an uncharged residue, the value listed will be -1.

## 3. MutMatrix:

These files record all mutations in a sequence in relation to the reference sequence. Only amino acid positions at which any sequence has a mutation from the reference sequence including deletions, insertions and point mutations are included. Therefore there are no columns for positions that are 100% conserved in these sequences – positions at which no mutations occur in any sequence.

The mutations are listed as the column header. For example “D30N” refers to a mutation at residue number 30 at which the wild-type amino acid, D, mutates to the amino acid N. A “1” means that the mutation is present; “0” means it is not present.

## 4. PNG:

These files list amino acid positions at which a potential N-linked glycosylation (PNG) site is present. The presence of a PNG is indicated by “1”; “0” means it’s absent. Sequences motifs of NxS/Tx, where X is any amino acid except for Proline are recognized by the program as a PNG.

## 5. PNG\_AA:

The amino acid sequence motif of each PNG detected and recorded in the “PNG” files is listed by position number.