

AminoTrackTM: Automating the Entry and Analysis of Mutations in Multiple Protein Sequences Using a Spreadsheet Format

Madhumita Mahalanabis
Department of Microbiology
University of Washington
Seattle, WA 98195

Jason Blue
Information Technology
Seattle Biomedical
Research Institute
Seattle, WA 98109

Nancy L. Haigwood
Departments of Microbiology and
Pathobiology, University of Washington
and Viral Vaccines, Seattle Biomedical
Research Institute Seattle, WA 98109

AminoTrackTM is a web based tool designed to increase the efficiency with which sequence data is recorded for further analysis. The main purpose of AminoTrackTM is to streamline the process and reduce human error in the identification of mutations present in multiple sequences compared to a reference sequence. Aligned protein sequences are entered in the web submission form and comma delimited files are generated in a zip file for loading into a spreadsheet. These files can be imported into any spreadsheet program that recognizes comma delimited files such as Microsoft Excel or SPSS. The sequences are analyzed for mutations in amino acids, charge changes, and potential N-linked glycosylation sites (PNG). The data may be viewed in a spreadsheet in a columnar binary format of "0" and "1" with one amino acid position per column. Currently this program is being used to identify mutations in viral proteins as these proteins evolve during infection.

Keywords: Protein Sequence, Program, Glycosylation, Mutation, Charge

1.0 Introduction

The most common method of analyzing protein or nucleotide mutation data is to perform a sequence alignment with a reference sequence and manually note mutations of interest by marking the alignment. This initial step is then followed by manually entering the noted mutations into a record such as a database, spreadsheet or written report for further data analysis (Figure 1). While there are many versatile programs to create alignments and a vast number of other programs to cover a multitude of sequence analysis functions, programs to record and compile mutation data in an alignment are not generally available. Thus while some or all of the steps in the mutation analysis process may be automated in larger research laboratories with on-site programmers to fill this gap, there is no freely available public source to allow a "wet-lab" researcher in an independent laboratory without access to programmers, to do such analyses in a completely automated fashion. The AminoTrackTM program was written to fulfill this basic need and provides a single environment for automating the recording and entry of vast amounts of mutation data. AminoTrackTM is a free online program for both Macintosh and PC with a graphical user interface complete with point and click buttons and text boxes familiar to the menu driven user. The program analyzes pre-aligned protein sequences and returns a variety of different outputs including amino acid mutations and mutations affecting the charge and PNG of a protein.

2.0 Materials and Methods

AminoTrackTM is a web based mutation analysis tool written in Visual Basic Net. The primary user interface is the main web page located at <http://apps.sbri.org/AminoTrack/>. The application consists of a web page to take user input and a back-end web service which handles most of the processing. The output files are created using simple string

comparison techniques built into the programming language. For example, the charges are calculated and the PNGs are identified for each sequence via standard methods. The sequences or strings are then compared and reported in separate files as mutations. Several of the files are primarily an automation of conversions, since a conversion is performed before the string comparison with standard conversion routines described below. The results are delivered in a zip archive file and contain comma-delimited text files usable in spreadsheet or statistics programs. The source code for AminoTrack™ is not freely available. AminoTrack™ is a trademark of Seattle Biomedical Research Institute and © Seattle Biomedical Research Institute, Seattle, Washington, 2006.

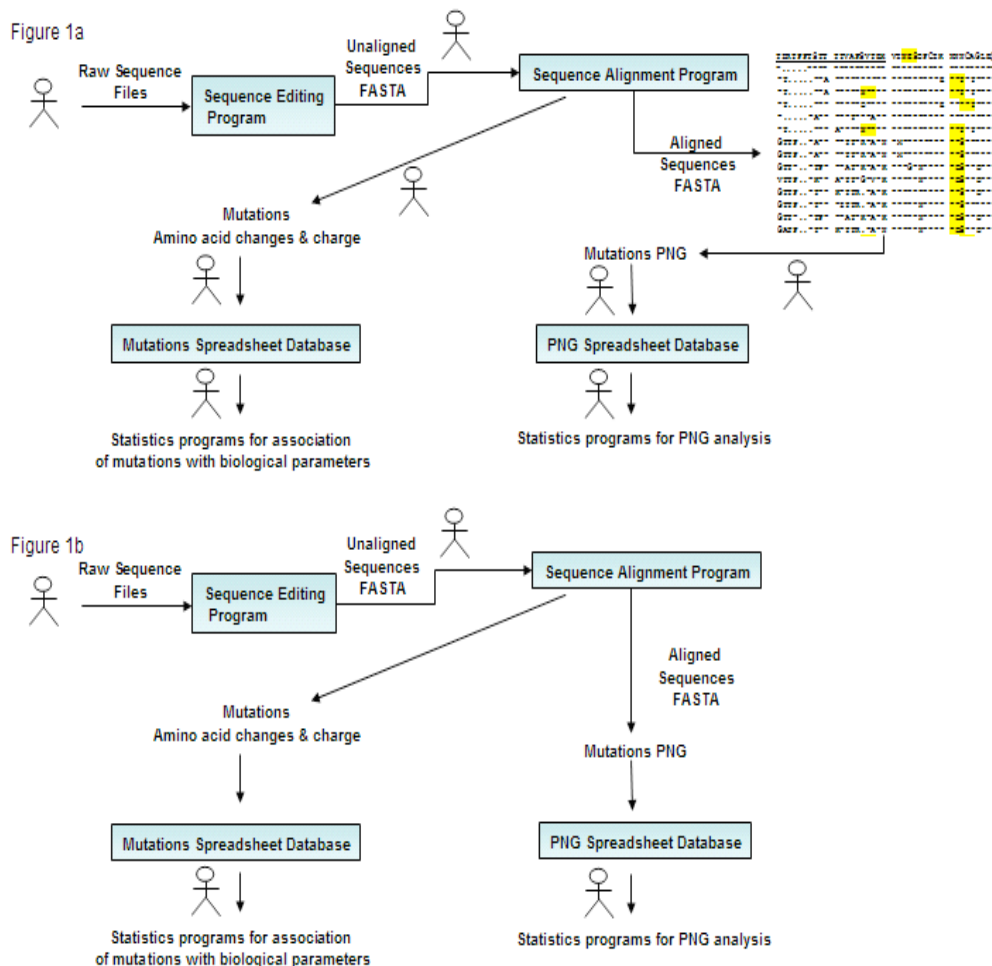


Figure 1. Top-level view of analyzing protein sequences for mutations.

The sequence analysis process (a) prior to the AminoTrack™ application contains numerous steps which the user must complete manually as indicated by the symbol of the human figure. This includes creating the alignment and marking mutations of interest on the alignment itself. Here the PNG changes from the reference sequence, the first sequence in the alignment, have been manually highlighted in the protein alignment. Similarly other amino acid mutations and changes in charged residues are typically recorded manually by first marking the alignment and then creating spreadsheet databases that are used for further data analyses. Many of the steps of the process are automated by AminoTrack™ (b), most importantly those of analyzing aligned protein sequences for mutations and creating spreadsheets for recording and analyzing the data.

3.0 Results

3.1 Algorithms to generate the mutation data

The web service is written in VB.Net and is the core of the application. When the request is submitted to the web service, the submitted sequence data is parsed using standard delimiters of #, >, and % to split the data into individual sequences. The first sequence in the data is considered the reference sequence and is used as a base for comparisons in other routines. The remaining sequences are maintained in an ordered list to be used by the various

output routines. Theoretically there is no limit on the number of sequences, but the maximum length of a string value or sequence length is 65,355.

There are several roles filled by the output routines. In the simplest case, a routine is used to reformat the sequence data into a labeled and delimited format usable in spreadsheets. Another routine compares the list of sequences to the reference sequence using a string comparison and outputs the changes. Several routines deal with amino acid charges and the changes in those charges between the reference sequence and each sequence in the list. In these routines, charge values are determined using standard formulas (amino acids D or E = -1, K or R = +1, else 0). While the charge routine simply displays the resulting charge values, the charge change routine calculates the amount of change from the reference sequence charge value to the compared sequence charge value (range -2 to +2). The AminoTrack™ web service also has a routine that identifies PNG locations. In this routine, each sequence is scanned for the accepted PNG patterns; NXS/TY (where X and Y can be any amino acid except P), and the main motif of N X S/T (where X can be any amino acid except P) [1-3]. If the routine finds a pattern match for either motif, the location of the start of the motif is identified in the output. An alternate output file from this routine shows the actual 4 character motif that triggered the PNG identification. All of the previously mentioned routines use the user-provided sequence offset identifier, starting amino acid position number, to provide a consistent position label across all reports. The final routine provided by the AminoTrack™ web service handles the creation of the mutation matrix for the submitted sequence data. This routine is slightly different in how it processes the data, as it makes multiple passes through the data. The first pass processes by sequence position instead of by submitted sequence order, comparing each sequence to the reference sequence to develop an internal list of changes or mutations found. Once the master list of mutations has been identified using standard notation of XyyyyZ (where X is the reference sequence amino acid, yyyy is offset position, and Z is compared sequence amino acid), the routine then processes each sequence in the submitted list noting the sequences that contain each identified mutation. The resulting file is a delimited list, using the mutations as labels, showing which sequences contain which mutations.

Due to the column count limits enforced by some spreadsheet programs, the application is designed to generate a file containing all data and also individual sub-files limited to 200 columns for easy import. The sub-files will contain the starting and ending offsets in their file names, while the complete file will contain the term “full” in its file name as an indicator.

3.2 Implementation

The current version of AminoTrack™ analyzes amino acid sequences that are already aligned with each other and a reference sequence. The user may use any alignment program such as CLUSTAL X for multiple sequence alignment that generates a text file of aligned sequences. The reference or wild-type sequence must be the first sequence in the sequence set and either the same length or longer than any of the other sequences for numbering purposes. Once the alignment quality is satisfactory to the user, and the sequences are in standard FASTA format, the data may be analyzed by AminoTrack™. To enter the aligned sequences, the user navigates to the AminoTrack™ website at <http://apps.sbri.org/AminoTrack/> and simply copies and pastes the aligned sequences into the sequence submission text box (Figure 2). The program also allows the user to specify the starting position number of the first amino acid in the sequences. This is typically the position number of the first residue in the reference sequence. AminoTrack™ numbers the amino acid positions according to the numbering rules for proteins of human and simian immunodeficiency viruses (HIV and SIV) [4, 5]. Numbering the amino acid residues allows the user to identify each amino acid and the region of the protein analyzed. Finally the user may enter a name for the resulting dataset. When the user submits the request, the page data is submitted to the web service for processing and receives back the resulting zip archive, which is then delivered to the user.

3.3 Mutation Analysis

3.3.1 Amino acid substitutions, insertions, and deletions

There are three file types that contain overall mutation data including all substitutions, insertions and deletions present in the sequences as compared to a reference sequence. The first is called 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 hyphen, while mutations are indicated by the letter of the mutant residue or a period in the case of a deletion (Figure 3a). The second file type is the MutMatrix file. 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 in Figure 3b, D33N refers to a

mutation at residue number 33 at which the wild-type amino acid, D, mutates to the amino acid N. A “1” means that the mutation is present; a blank cell means it is not present. The third file is called AASeq and contains the full length of each amino acid sequence submitted. The file presents the sequences by creating an individual column for each position that lists the identity of the amino acid for all sequences (not shown). This is essentially a view of the alignment in a spreadsheet format and can be referred to as the user compares mutations amongst the different sequences using the other file types.

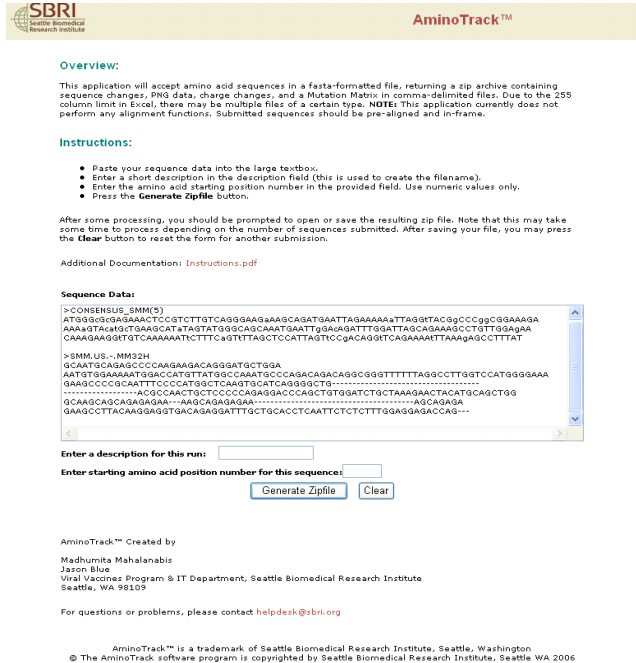


Figure 2. Screenshot of the AminoTrack™ data submission page.

The main web page for the application is shown here. The user enters the FASTA file of aligned protein sequences in the sequence data text box as shown. The reference sequence must be the first sequence in the alignment followed by all other sequences. A name for the resulting zip archive file may be entered in the text box asking for a run description. If the starting position of the protein sequences is different than one, it may be entered in the third text box as indicated. Once this information is entered, the user may analyze the data and create spreadsheets by selecting the “Generate Zipfile” button.

3.3.2 Charge

Two file types also pertain to data about amino acid charge. The first type is called Charges. The charge for each charged residue in the sequence 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 Microsoft Excel (Microsoft Corporation, Redmond, WA), for example, for each sequence. An example of the output is shown in Figure 3c. All sequences have amino acid K at position 31 and amino acid E at position 17. Thus the values listed are +1 at position 31 and -1 for position 17.

The second charge output has the file name ChargeChanges. This is similar to the charge output above except that the individual charge values for each residue are not given. At positions that have a mutation affecting charge, the file lists the resulting change in charge from the reference sequence. For example in Figure 3d, since there is a K (reference sequence HXB2) to E (mutant sequence SHIVSF162P3) mutation at position 51, the charge changes from 1 (K) to -1 (E). The value listed at position 51 is -2 for the mutant sequence, since there is a net change of -2. None of the other sequences have a mutation at this position (see figure 3a); therefore the cell is blank, since there was no net change in charge. Similarly for an E (-1) to K (1) change, the value listed would be +2.

3.3.3 N-linked Glycosylation

Files by the name of PNG list amino acid positions at which a PNG site is present. The presence of a PNG is indicated by “1”; a blank cell indicates its absence. This spreadsheet format allows the user to calculate the frequency of PNG sites at each position in the sequence set. The data can be graphed directly in Excel or imported into other data analysis and graphing programs. The sequence motif defining a PNG site is NXS/TY, where N, X and Y, S, and T are the amino acids asparagine, any amino acid except for proline, serine, and threonine, respectively. A second file type called PNG_AA, lists the specific amino acid sequence of the PNGs at each position in each sequence (Figure 3e and 3f).

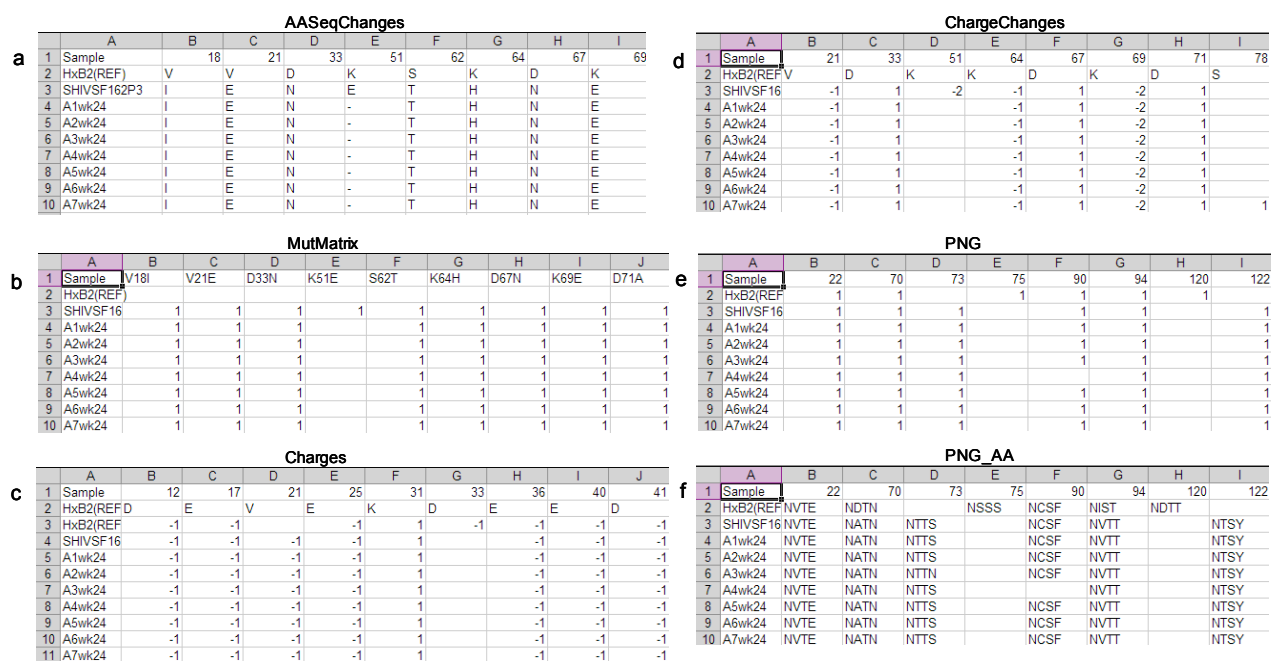


Figure 3. Screenshots of the generated mutation data in spreadsheets.

All data files generated by the application are comma delimited files which may be viewed as spreadsheets in programs such as Microsoft Excel or statistics programs such as SPSS. Here the six main data file outputs are shown in Microsoft Excel spreadsheets. In each case the data is in a column format with one amino acid position per column and the reference sequence in the first row. The sequence order is the same as that of the submitted sequence FASTA file. Only positions with relevant data are included in these files, such that positions without mutations are excluded from the reports. The AASeqChanges file (a) shows each amino acid sequence change from the reference and the position number. Sequences that do not have changes at the position have a hyphen. The MutMatrix files (b) list mutations in the column header and the sequences with these mutations by entering a value of "1." In the Charges files (c) the charge value of each charged residue in each sequence is given. A blank cell here indicates that no charged residue is present at that position. Any change in the charge value at a given position is recorded in the ChargeChanges files (d) such that the net gain or loss of charge is listed at each position where a change occurs. The PNG files (e) list positions at which a PNG site is present in any sequence by entering the value of "1." The actual amino acid sequence motif of each PNG is recorded in the PNG_AA file (f).

4.0 Discussion

Many of the functions of AminoTrack™ are automation routines designed to easily generate delimited data files from sequence data. AminoTrack™ provides reports displaying charge calculations, charge changes, amino acid sequence changes, PNG locations, and a mutation matrix from a submitted FASTA format list of sequences. Standard string comparison techniques are used to generate most reports. In the case of charge calculations and PNG locations, standard conversion rules are applied before comparisons and report generation.

This application was initially created for the purpose of analyzing mutations that occur in the Envelope (Env) protein of HIV and SIV as it evolves in response to selective pressures from the host's immune responses during infection. HIV replicates with extensive genetic variation within individual hosts, particularly in the hypervariable regions of Env [6]. Such rapid evolution is the result of a combination of factors. First, the virus experiences a high rate of mutation, with reverse transcriptase enzyme, replicating the virus genome, making around 0.2 errors per genome during each replication cycle [7], and further errors occur during transcription from DNA by RNA Pol II polymerase. Second, HIV replicates at a rapid rate with a viral generation time of ~2.5 days and produces approximately 10^{10} – 10^{12} new virions each day [8]. Finally, frequent recombination and natural selection in the host further elevate its rate of evolutionary change. There is strong evidence that positive selection is the driving force of evolution within a host.

HIV accrues mutations that allow it to evade immune responses, particularly in Env. Host immune selection pressure could be generated by antibodies that block virus infection of target cells called neutralizing antibodies

(NAbs). NAb responses are generated early in infection at first to the early Env variants and then to subsequent variants. The antibody responses to these variants exert a selective pressure that drives continuous evolution of neutralization escape mutants [9]. One known example of mutations that are important with regard to virus neutralization is the mutation of sequence that alters the placement and number of sugar moieties, N-linked glycans, on Env. The surface unit of Envelope of HIV contains on average 25 sites for potential N-linked glycosylation. These glycans are complex branched-chain carbohydrate structures that in total account for approximately 50% of the molecular mass of the glycoprotein. Mutations in the motif in Env may result in the addition and rearrangement of N-linked glycans that shield the virus from antibody mediated neutralization. Although alterations in this “glycan shield” occur during the course of infection, allowing the virus to survive by escaping antibody recognition [10], there are constraints on the fluidity of the glycan shield such that mutations altering glycosylation occur in limited regions of Env [11].

AminoTrack™ has a number of advantages and limitations that are a result of initial decisions made to achieve the best balance of achieving the data analysis and file types described and programming and computing issues. Several of the advantages include the portability of the program on both PC and Macintosh, accessibility on the web, web based design increasing ease of technical support and maintenance, and generation of comma delimited files importable in many common programs. Limitations include the analysis of protein, but not nucleotide sequences and the size limit of data importable into a single Excel spreadsheet.

The decision to build AminoTrack™ as a web based application was made early on, and was influenced primarily by the desire to have some form of centralized control of the application code. Adding new features or fixing bugs that may be discovered is simplified for web based applications which tend to be centrally located. This approach also leads to more consistent user results. There are no version discrepancies which may cause confusion in the user base, since new features will be available to all users and not only users who may have downloaded the latest version. The downside to this approach is that there are server and internet bandwidth requirements that do not exist with a stand-alone application that is downloaded on individual computers. Also, users must have internet access in order to use the application. Considering the prevalence of internet connections in homes and workplaces today, it was determined that the benefits of providing a consistent usage experience throughout the applications life outweighed the negatives of the web application design.

Another consideration was in delivering file types that could be used in the majority of spreadsheet and statistics programs. The comma delimited format was chosen for this purpose for all files, as it is considered to be the most widely used and supported import format available. We anticipated that Microsoft Excel would be the most commonly available and used program for viewing the data and decided to make the files importable in Excel. In doing so, the limitation of a 255 column count limit is introduced by the Excel software. Thus if there are more than 255 mutations in the sequence, the column limit is exceeded. Importing more than 255 columns causes Excel to give an error and fail data import. This problem could not be completely solved by simply transposing the data such that each column represents a sequence and each row of the spreadsheet represents an amino acid position. This solution would not be viable for datasets with more than 255 sequences as in the case of clinical laboratories, for example, that must maintain patient databases with well over 255 patient sequences. Therefore the original format as shown in Figure 3 was maintained. To solve the issue and maintain the individuality of the data at each position (to use the formula functions in Excel) individual columns were preserved instead of combining values into single columns. As a result, the reports are generated in divisions to fit Excel’s column limits while maintaining a “full” file for use in other programs without this limitation. Therefore the program was designed to generate sub-files limited to 200 columns for easy import into Excel, along with a master file containing all data usable in other spreadsheet programs or statistical packages. These sub-files will contain the starting and ending position numbers or offsets in their file names, while the complete file will contain the term “full” in its file name as an indicator.

Finally, the ability to convert submitted nucleotide sequences to their resulting amino acid sequences prior to running the normal reporting routines is not supported in AminoTrack™. This would give the user the ability to submit either a series of nucleotide sequences or a series of amino acid sequences and end up with results in a similar format. While a straight chart-based conversion is trivial to perform in code, we realized that there are numerous steps including alignment that are outside of the original intended scope of AminoTrack™. There are numerous other programs in existence to handle these steps already, and the user may easily convert nucleotide sequences into amino acid sequences.

Beyond the HIV studies mentioned above, the task of tracking mutations in any biological system in virtually any protein will be simplified by using programs such as AminoTrack™ that eliminate the tedious, time-consuming and error-prone task of noting mutations by hand. While there are innumerable programs to perform tasks of sequence analysis covering a vast array of different functions (see the Biological Information Resources at the University of Washington for a small sampling at <http://courses.washington.edu/bioinfo/BIR/>), no program was available that fulfilled the functions of AminoTrack™. While the program N-glycosite at the Los Alamos National Laboratory's HIV Sequence Database, analyzes protein alignments for PNG sites and PNG motifs, it returns only a pre-made graphical output for PNG positions and frequencies [12]. The data is not compiled into a spreadsheet format with the PNG data individually for each sequence at each position for the user's records to analyze or present in other formats. Therefore the main benefits of AminoTrack™ are the automation of the process and the transferring data to a spreadsheet format to create a record of mutations for further analysis in commonly used spreadsheets such as Microsoft Excel or statistical programs.

References

- [1] R. D. Marshall, "The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins," *Biochem Soc Symp*, pp. 17-26, 1974.
- [2] L. Kasturi, H. Chen, and S. H. Shakin-Eshleman, "Regulation of N-linked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-Xaa-Ser/Thr sequons that are poor oligosaccharide acceptors," *Biochem J*, vol. 323 (Pt 2), pp. 415-9, 1997.
- [3] J. L. Mellquist, L. Kasturi, S. L. Spitalnik, and S. H. Shakin-Eshleman, "The amino acid following an asn-X-Ser/Thr sequon is an important determinant of N-linked core glycosylation efficiency," *Biochemistry*, vol. 37, pp. 6833-7, 1998.
- [4] B. T. Korber, B. F. Foley, C. I. Kuiken, S. K. Pillai, and J. G. Sodroski, "Numbering Positions in HIV Relative to HXB2CG," in *Human Retroviruses and AIDS. Report LA-UR 99-1704*, B. T. Korber et. al., Ed. Los Alamos, NM: Los Alamos National Laboratory, 1998, pp. III-102&endash;III-111.
- [5] M. J. Calef C, O'Connor DH, Watkins DI, Korber BT, "Numbering Positions in SIV Relative to SIVMM239," in *HIV Sequence Compendium*, F. B. Kuiken C, Hahn B, Marx P, McCutchan F, Mellors JW, Wolinsky S, Korber B., Ed. Los Alamos: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2001, pp. 171-181.
- [6] E. Holmes, L. Zhang, P. Simmonds, C. Ludlam, and A. Brown, "Convergent and Divergent Sequence Evolution in the Surface Envelope Glycoprotein of Human Immunodeficiency Virus Type 1 within a Single Infected Patient," *PNAS*, vol. 89, pp. 4835-4839, 1992.
- [7] B. D. Preston, B. J. Poiesz, and L. A. Loeb, "Fidelity of HIV-1 reverse transcriptase," *Science*, vol. 242, pp. 1168-71, 1988.
- [8] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho, "HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time," *Science*, vol. 271, pp. 1582-6, 1996.
- [9] D. D. Richman, T. Wrin, S. J. Little, and C. J. Petropoulos, "Rapid evolution of the neutralizing antibody response to HIV type 1 infection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 4144-4149, 2003.
- [10] X. P. Wei, J. M. Decker, S. Y. Wang, H. X. Hui, J. C. Kappes, X. Y. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw, "Antibody neutralization and escape by HIV-1," *Nature*, vol. 422, pp. 307-312, 2003.
- [11] W. M. Blay, S. Gnanakaran, B. Foley, N. A. Doria-Rose, B. T. Korber, and N. L. Haigwood, "Consistent patterns of change during the divergence of human immunodeficiency virus type 1 envelope from that of the inoculated virus in simian/human immunodeficiency virus-infected macaques," *J Virol*, vol. 80, pp. 999-1014, 2006.
- [12] M. Zhang, B. Gaschen, W. Blay, B. Foley, N. Haigwood, C. Kuiken, and B. Korber, "Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin," *Glycobiology*, vol. 14, pp. 1229-46. Epub 2004 Jun 02., 2004.