

CView: A network based tool for enhanced alignment visualization

Raquel Linheiro^{\$,1}, Stephen Sabatino^{\$,1,3}, Diana Lobo^{\$,1,2,3} and John Archer^{*,1,3}

¹ CIBIO/InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus de Vairão, Portugal.

² Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Portugal

³ BIOPOLIS, Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, Portugal.

* john.archer@cibio.up.pt
\$ contributed equally



ABSTRACT

To date basic visualization of sequence alignments have largely focused on displaying per-site columns of nucleotide, or amino acid, residues along with associated frequency summarizations. The persistence of this tendency to the recent tools designed for viewing mapped read data indicates that such a perspective not only provides a reliable visualization of per-site alterations, but also offers implicit reassurance to the end-user in relation to data accessibility. However, the initial insight gained is limited, something that is especially true when viewing alignments consisting of many sequences representing differing factors such as location, date and subtype. A basic alignment viewer can have potential to increase initial insight through visual enhancement, whilst not delving into the realms of complex sequence analysis. We present CView, a visualizer that expands on the per-site representation of residues through the incorporation of a dynamic network that is based on the summarization of diversity present across different regions of the alignment. Within the network, nodes are based on the clustering of sequence fragments that span windows placed consecutively along the alignment. Edges are placed between nodes of neighbouring windows where they share sequence identification(s), i.e. different regions of the same sequence(s). Thus, if a node is selected on the network, then the relationship that sequences passing through that node have to other regions of diversity within the alignment can be observed through path tracing. In addition to augmenting visual insight, CView provides export features including variant summarization, per-site residue and kmer frequencies, consensus sequence, alignment dissection as well as clustering; each useful across a range of research areas. The software has been designed to be user friendly, intuitive and interactive. It is open source and an executable jar, source code, quick start, usage tutorial and test data are available (under the GNU General Public License) from <https://sourceforge.net/projects/cview/>.

PLOS ONE Linheiro, R., Sabatino, S., Lobo, D., & Archer, J. (2022). CView: A network based tool for enhanced alignment visualization. PLOS ONE, 17(6), e0259726.

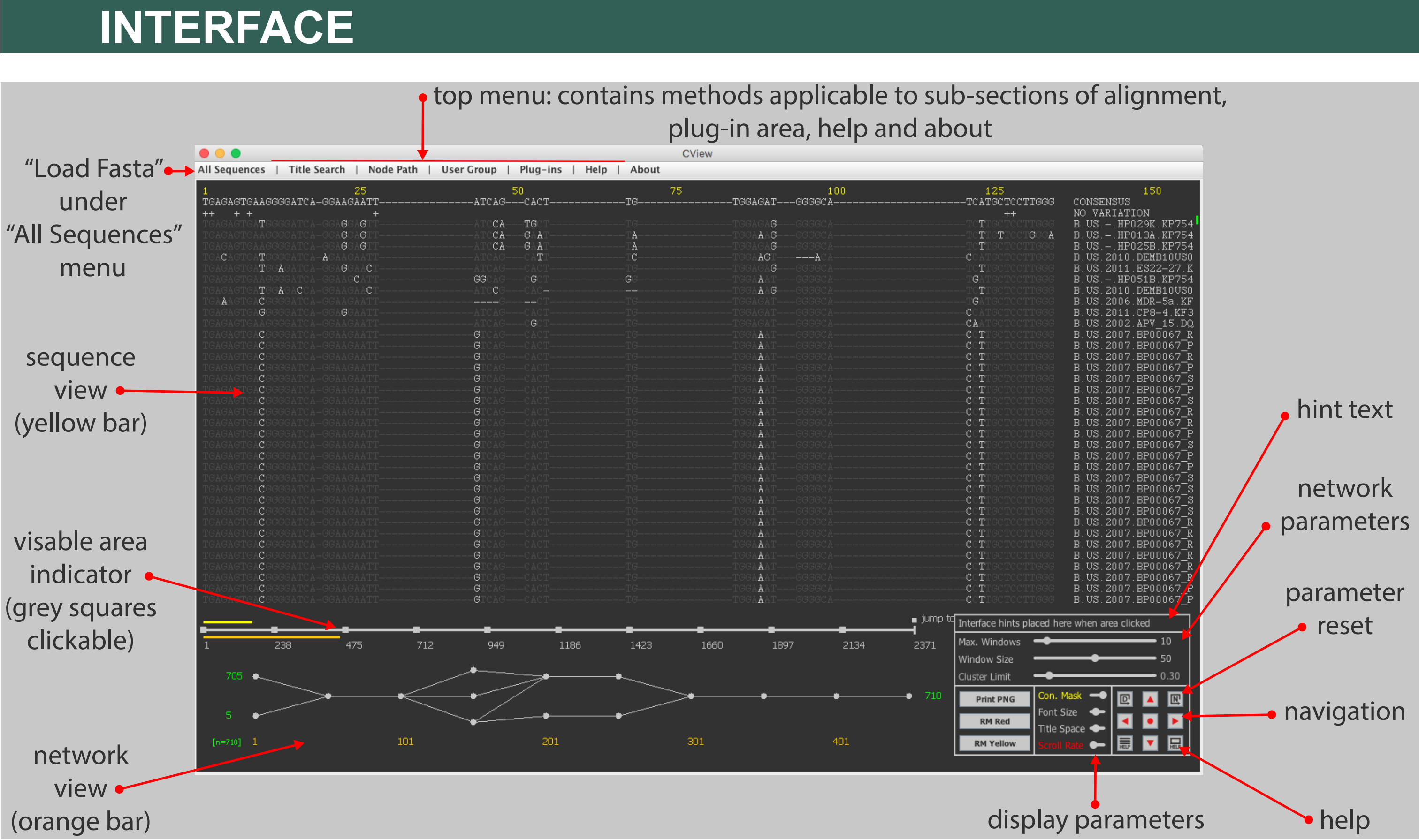


Figure 1: CView interface. The four main areas of the CView interface are depicted. These are sequence view, network view, control panel and the top menu. The yellow numbers on the top indicate the sites of the alignment that are currently in view. These correspond to the yellow bar on the top of the location indicator. The orange numbers along the bottom indicate the locations of the windows that nodes within the network are dependent on. These window locations correspond to the area that the orange bar located under the location indicator covers. Grey dots indicate selectable nodes within windows. The squares along the location indicator can also be selected in order to jump directly to the indicated co-ordinates. The red text around the outside of the interface summarizes the main features.

DESIGN AND IMPLEMENTATION

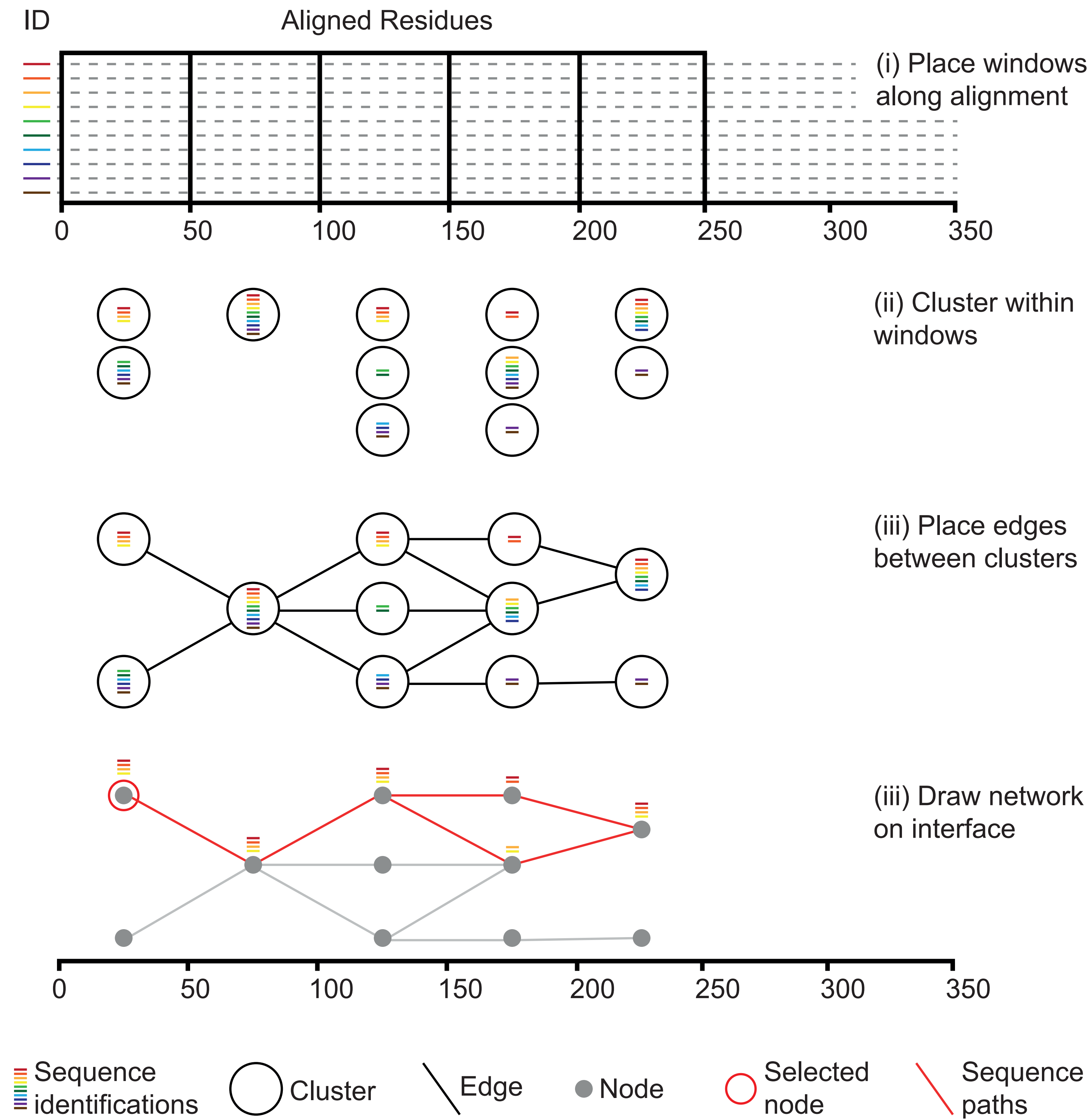


Figure 2: Network construction. Coloured bars indicate unique sequence id's relative to the corresponding sequences (dotted lines). Within each window the identification representing each full sequence are associated with individual sequence fragments spanning that window (i), and fragments within windows are clustered (ii). Edges are placed between neighbouring clusters where they share one or more sequence identification, i.e. differing regions of the same sequence (iii). Clusters are represented visually on the network by grey dots. If a single cluster is selected the paths of all sequences passing through in relation to all other clusters (red lines) can be traced (vi).

(i) Label clusters and identify required edges using shared sequence id's between clusters

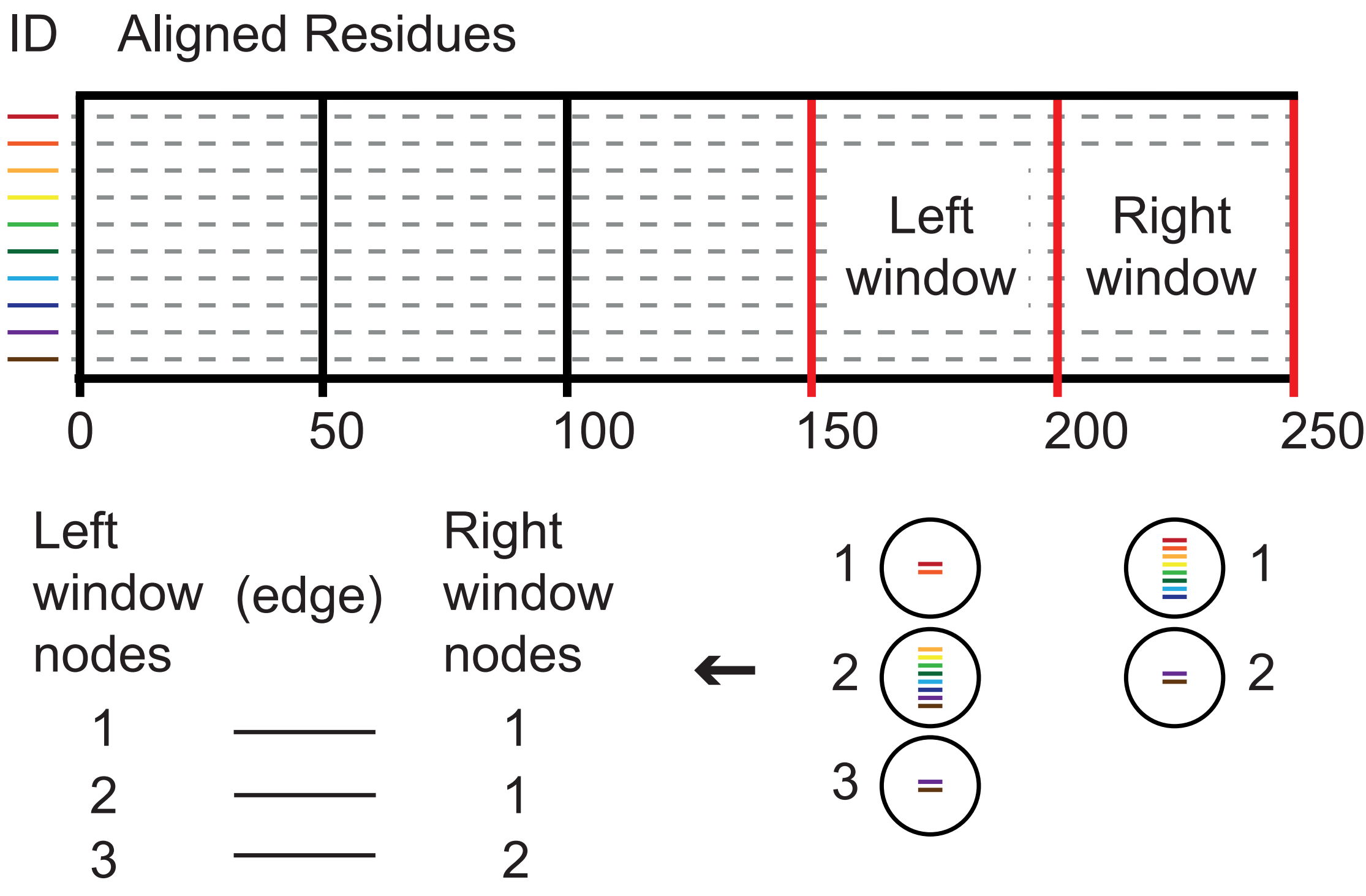


Figure 3: Minimization of edge crossovers between nodes of the two right most windows of the alignment. This process is repeated until the left most window (anchored on site 1) is reached. Clusters within the two windows are labelled with integers and required edges, based on the sequence ids (coloured bars), are listed (i). All order permutations of the current left window are identified and for each permutation the required edges are placed relative to the constant cluster order of the right window (ii). Crossovers are then counted (red numbers). Of the permutations that produce the minimum number of crossovers a random one is selected for graphical node layout order.

CONCLUSION

Basic alignment visualization should have the potential to increase the level of initial insight within sequence datasets whilst not delving into the realms of more complex sequence analysis. CView is a tool that allows the user to interactively explore sequence alignments with the aid of a dynamic network that summarizes the diversity present within regions of the alignment not currently in-view. CView provides a range of export features that can be applied to the entire alignment, to a specified region of the alignment, or to a specified region in conjunction with a specified subset of sequences. Such export features include variant summarization, per-site residue and kmer frequency matrixes, clustering, pairwise-distance matrixes as well as consensus sequence generation. The exact usage scenario in which CView can be applied is dependent on the requirements, insight and background of the individual user.

RESULTS

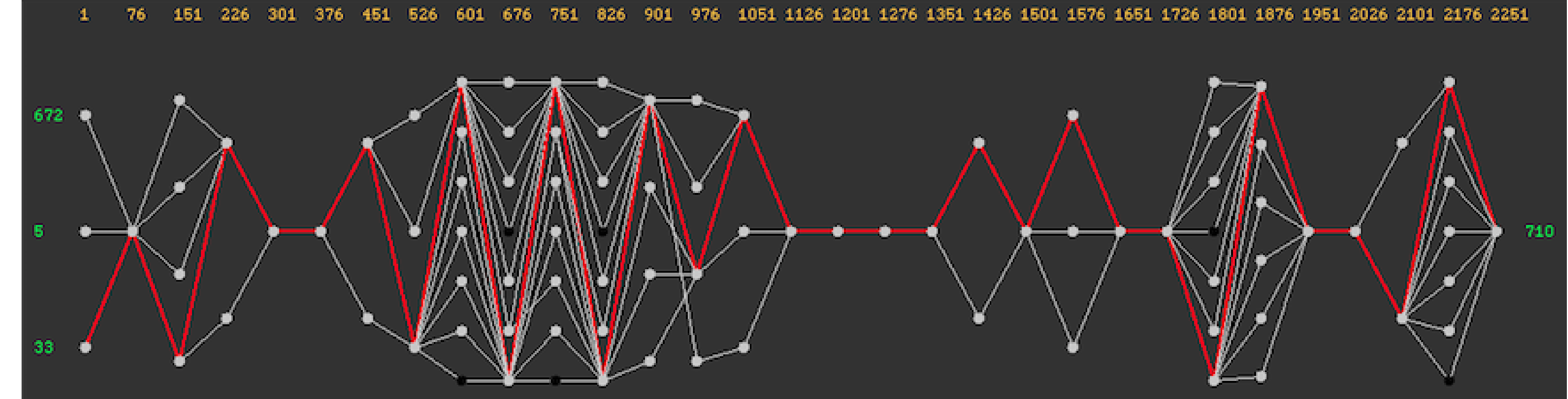


Figure 4: Network based diversity across the full gp120 case-study alignment. Thirty non-overlapping windows of length 75 nt fitted along the length of the alignment. Details of the alignment used are available in Linheiro *et al.*, (2022) [1]. The clustering threshold within each was 0.2. The orange numbers along the top indicate window locations, whilst the green numbers on each side indicate the number of sequences observed within nodes (clusters) at the ends of the network. Grey dots are clickable nodes, that when clicked will highlight the paths (in red) of the sequences passing through across the remainder of the network. Here the highlighted paths are a result of clicking the bottom left most node (containing 33 sequences). Information from sequences highlighted in this manner can be extracted using the sub-options of the "Node Path" top-level menu within the software as indicated within the demonstration video [2], and under the software help menus. This network was printed using the "Print PNG" button located within the control panel of the software.

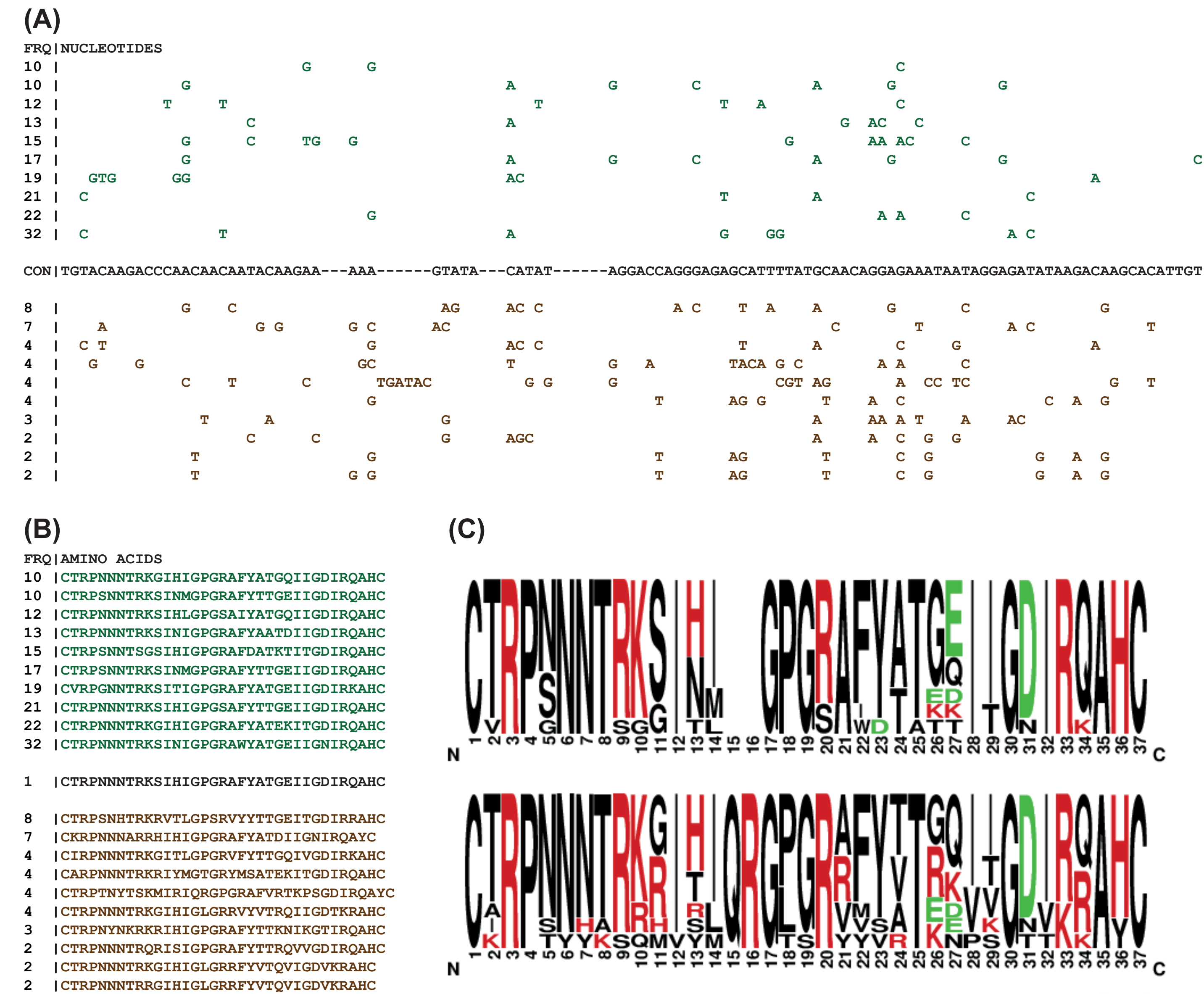


Figure 5: Summarization of variation present within the V3 loop. For the alignment used in figure 4: (A) Green residues represent non-consensus residues from the ten most frequent variants associated with the CCR5-using phenotype. Brown represents those of the CXCR4-using phenotype. The consensus sequence (black) is shown. (B) Translations of the most frequent ten variants from each phenotype. (C) Sequence logos summarizing these translations. The top logo is from represents the CCR5-using sequences whilst the bottom represents the CXCR4-using ones.

REFERENCES

1. Linheiro, R., Sabatino, S., Lobo, D., & Archer, J. (2022). CView: A network based tool for enhanced alignment visualization. PLOS ONE, 17(6), e0259726.

2. Linheiro, Raquel, Lobo, Diana, Sabatino, Stephen, & Archer, John. (2022, May 3). CView: tutorial 1 - overview (movie and script). Zenodo. <https://doi.org/10.5281/zenodo.6514787>.

FUNDING

The Portuguese Foundation for Science and Technology, FCT, projects NORTE-01-0246-FEDER-000063, PTDC/BIA-E-VL/29115/2017, POCI-01-0145-FEDER-029115 and PD/BD/132403/2017.

