

## CDTree: a tool to analyze and annotate protein subfamily hierarchies

Chunlei Liu, David Hurwitz, Christopher Lanczycki, Aron Marchler-Bauer, Vahan Simonyan, Stephen H. Bryant  
*Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland*

### Introduction:

Conserved protein domains are units of molecular evolution, as they tend to occur in various contexts, and typically have undergone gene duplication events and functional diversification. For the most part such domains are compact and distinct units of protein three-dimensional structure and have well characterized molecular function. Curated collections of protein domain models have been used successfully in the computational annotation of protein sequences, and very often the presence of conserved domain signatures is all the functional annotation available for genomic sequences. The Conserved Domain Database (CDD) project is a curation effort in NCBI to re-organize redundant collections of domain and protein models from different sources such as the Pfam, SMART, and COG. During the course of curating the CDD database, it becomes increasingly clear that it is necessary to combine obviously related models into hierarchies, where each hierarchy spans a domain (or protein) super-family. Such a hierarchy would enable us to capture both functional or structural features specific to individual families and the commonality shared by all the sub-families with a hierarchy.

Performing hierarchical analysis of these families requires integration of knowledge from various sources such as structure-guided multiple sequence alignments, phylogenetic analysis, the taxonomic distribution of aligned sequences, domain architectures, and the published literature. While algorithms or databases have been developed to allow for these analyses, no application has been reported to effectively integrate all the necessary analysis within one user-friendly graphical user interface (GUI) environment.

We have developed the CDTree software to fill this void. CDTree is an interactive graphical application designed to discover and create hierarchical relationships among domain families in a consistent, coherent fashion. CDTree has an interface to PSI-BLAST search using the existing search models of a domain family to scan the protein sequence database, and explore the resulting sets of consistently aligned sequences. Consistent and extensible GUI-based interfaces allow one to easily perform phylogenetic analysis, profile-recognition studies, and domain architecture examination. Integration with the existing application Cn3D empowers the application with structure-guided alignment editing capabilities. In order to easily correlate different sources of information, proteins underlying features discovered in each type of analysis can be highlighted and the highlights are also reflected in all other concurrent analysis on the same domain.

### Main Features:

The main window of the CDTree application presents a hierarchical view of subfamilies within one or more domain families and provides visual cues to select families of interest for further analysis. The main window also serves as the “launch pad” for other types of detailed analysis focused on the selected families. Each type of analysis – phylogenetic trees, domain architectures, profile-recognition, and taxonomy information – is presented in a separate window (called viewer in the application). Information highlighted in one viewer is automatically conveyed to the other viewers for easy data tracking between viewers. There are numerous useful features in the software to manipulate the alignments with a domain hierarchy by moving and re-indexing alignments with subfamilies, create a new subfamily, detect alignment inconsistencies in the hierarchy, and remove redundant sequences from a family. However, here we only list the following key viewers in discovering and declaring the subfamilies with a domain:

Sequence Tree: This viewer calculates and displays a phylogenetic tree of sequences in one or more protein families. Tree branches can be selected and thus investigated as a candidate as a new subfamily in combination with other information on the selected branch. To generate the tree, a distance matrix is first calculated by using one the similarity measures: percent-identity of aligned residues, scores of the aligned residues, BLAST scores of the aligned regions and BLAST scores of the whole sequences. The phylogenetic tree is derived by applying single-linkage or neighbor-joining clustering to the distance matrix.

Taxonomy Tree: This viewer is a browsable taxonomy tree to show the distribution of proteins in selected protein family across the taxonomy span. The source of taxonomy information is the NCBI taxonomy server.

Cross Hit: This viewer compares the ability of a domain family’s profile or PSSM to recognize the sequences within the family versus its ability to recognize the sequences within the sibling subfamilies. The e-values of PSI-BLAST are used as the measure for the recognition capability and the results are displayed as interactive bar charts and pie charts. This analysis helps to reveal proteins that more suited to be classified into a different subfamily.

Domain Architecture: This viewer makes use of the CDart database, group the sequences within one or more protein families based their domain architecture, and, display the clustering in a graphical hierarchical style.

CD-Update: This feature can be used to recruit new sequences into a domain family. It present a interface to perform PSI-BLAST searches against GenBank using the profile of the family. The new sequences found from the searches are processed and added to the family. Frequently, when new sequences are added to a family, a new subfamily emerges from it.

### Platform and Availability:

CDTree has been successfully used in curating the CDD database. We plan to make it downloadable from CDD web page and use it a helper application to further explore domain hierarchies. The application has been developed in C++ using the platform-independent API, wxWidgets (<http://www.wxwindows.org/>) for GUI development and relies on the NCBI C++ toolkit ([http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP\\_DOC/](http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/)) to encode and decode the biological data sets. Like its companion structure-based alignment editor, Cn3D, CDTree is designed to run on Windows, Mac and UNIX. The Windows version is now available and porting to other platforms is under way. To request the program, send email to [cliu@ncbi.nlm.nih.gov](mailto:cliu@ncbi.nlm.nih.gov).