white paper

Genchek™

A Comprehensive Sequence Analysis Package

Introduction

Study of DNA, RNA and amino acids has been a part of research in life sciences for more than three decades. Analysis of sequences of these molecules is fast becoming integral and indispensable to life sciences research, whether it is population genetics, evolution, plant breeding or pharmacogenomics.

Sequence data is linear (one dimensional), however it is this linear sequence of molecules such as DNA (thus RNA, and eventually amino acids) that manifest in external variation or a phenotype of interest to biologists, irrespective of the species or cellular mechanisms they study therein. A typical sequence analysis process may include base composition analysis, sequence quality analysis, sequence trimming for vector contamination, search for similar sequences in a public or proprietary database, analysis of sequence for putative coding regions and estimation of amino acid that may be coded from a string of DNA. Each of these processes have nuances of their own, wherein a researcher must be able to define individual parameters to evaluate data.

A sequence analysis software's primary objective is to filter, organize and analyze nucleic acid and amino acid data to generate tangible information. Genchek™ has been conceived, designed and developed by researchers keeping in mind the requirements of a sequence analysis research lab.

The Problem

With advent of automation and progressive sophistication of diagnostic techniques, the amount of data generated is far greater than in previous years. Managing such a magnitude and complexity of tasks and the information that is integral to it apart from core process research, i.e., to analyze and think, can be a daunting task. Typically, routine tasks and processes consume a large part of a researcher's workday. If these tasks can be organized and the work flow synchronized to add efficiency, then it will provide a researcher more time and resources to think and analyze.

Genchek™ : One-stop Platform for Comprehensive Sequence Analysis

Genchek™ is a comprehensive, platform independent, sequence analysis package based on client-server architecture and built on a relational database system. Genchek™ facilitates analysis of genomic, proteomic and SNP (Single Nucleotide Polymorphism) data. The need for such sequence analysis tools stems from requirement to collect, handle, manipulate, analyze, and annotate sequence data.

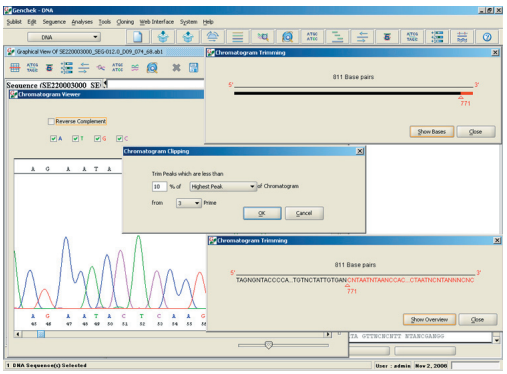
Genchek™ allows analysts to perform different operations, right from capture of a chromatogram to final annotation of a sequence. It supports all critical steps of a sequence analysis experiment. It can help a researcher to:

- Capture, organize and analyze biological sequence data easily
- Manipulate raw data for better analysis and informative output
- Provide easy and timely access to pertinent information
- Report comprehensive and easy-to-read results

What Genchek™ Extends

Electropherogram Viewer

This tool allows import of an electropherogram (chromatogram) from a sequencing machine. A user can trace all standard electropherograms (for example *.abi., *.scf, *.abd) using this tool. It also allows a researcher to rescale an image, edit a sequence, clip noisy peaks and convert electropherogram data to text format.

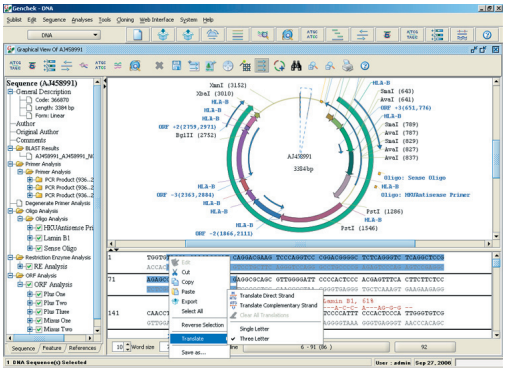


Sequence Editor and Graphical Viewer

The user can either import a sequence from a local file, which can be in GenBank, EMBL, FASTA format or a plain text file with just the sequence or select from the data browser. This module enables a user to modify the selected sequence

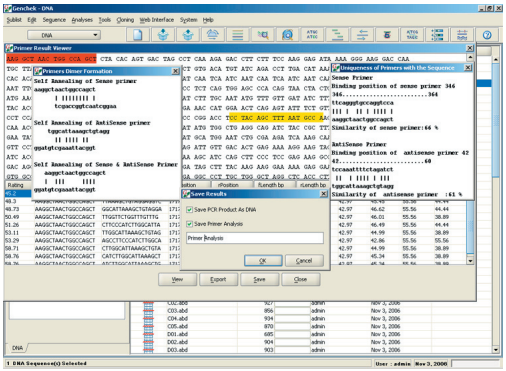
Ocimum Biosolutions
...enabling R&D™
www.ocimumbio.com

or create a sequence by simply keying in bases of a sequence or by copying a sequence from a text editor and pasting it to sequence editor. It accepts only valid bases and standard ambiguities for nucleotides and amino acids. The Graphical Viewer represents the sequence in both graphical and text formats, helping user to perform various analyses (such as ORF Analysis, RE Analysis, Oligo Analysis, etc.) and track results obtained. This also allows users to add annotation for sequences.



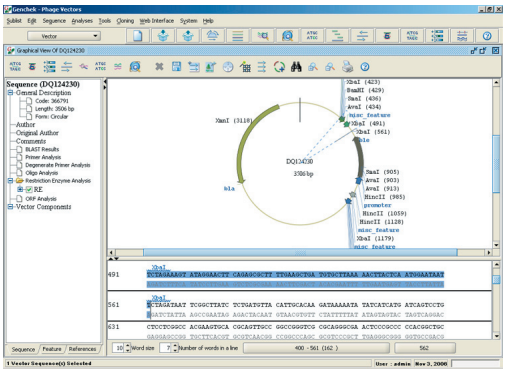
Primer Design

The key to PCR lies in the design of two oligonucleotide primers. This tool helps in designing the optimized primer for any nucleotide sequence. All possible primers are represented in a tabular manner and selecting any row highlights the respective primer in the text viewer. User has the choice to select parameters like primer length, %GC content, 3' sequence etc. You can save the designed primers, PCR products and also save these primers as oligos and DNA patterns. Degenerate primer design to amplify homologous sequences and primer design for multiple sequences are also provided.



Cloning Module

Cloning module in Genchek™ helps in designing new vectors with RE-based cloning and Gateway cloning supported by graphical interface of sequences. It supports creation of a vector molecule from fragments that are completely defined and made compatible by the user. Gateway cloning module facilitates LR and BP reactions of Gateway cloning protocols and generation of B-cassettes through PCR amplification.



Contig Assembly

Contig assembly tool is used to arrange or assemble several sequences into single consensus sequence of higher length. It allows overlap region length selection, fragment inversion check and identification of consensus regions. Contig results and contig map can be saved and exported to a desired location.

Sequence Alignment Module

Using Genchek™ Sequence alignment modules, user can perform **Global** and **Local** alignments on pair of sequences, **Dot Plots** for locating possible direct and inverted repeats. User can also align multiple sequences, study phylogenetic relationships between sequences and find out conserved regions **in** a set of sequences using Multiple Alignment Editor in **Multiple Sequence Alignment** tool.

Major functions of Multiple Alignment Editor include :

Import and Export of Alignments and Trees: Alignments can be imported and exported in different formats (Fasta, PFAM, MSF, Clustal, BLC, PIR). Associated trees can be imported and exported in newick format.

Editing: Gaps can be inserted/deleted using the mouse or keyboard. Group editing (insertion,

Ocimum Biosolutions
...enabling R&D™
www.ocimumbio.com

deletion of gaps in groups of sequences) is also possible in the Multiple Alignment Editor.

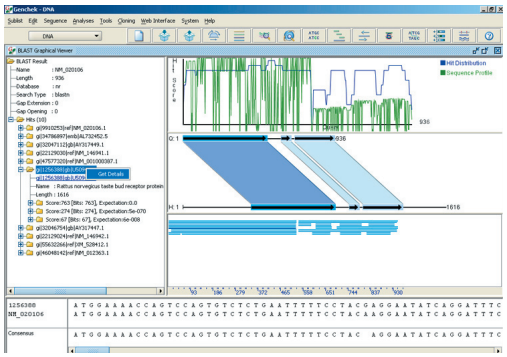
Analysis: Alignments can be sorted on these options, i.e., by name, by tree order, by percent identity, by group. UPGMA and NJ trees are calculated and drawn based on percent identity distances. Sequence clustering can be performed using principal component analysis. Pairwise alignment of selected sequences is also possible.

Annotation: To color alignments or groups, user predefined or custom color schemes are available. Sequence feature retrieval and display option is also possible on the alignment.

Publishing: Align ment can be printed with colors and annotations. Output alignment are exported as portable network graphics image (PNG) and encapsulated postscript file (EPS).

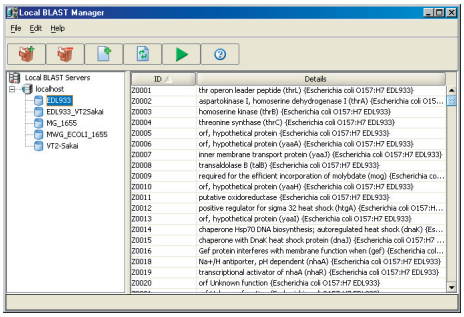
BLAST Suite

Genchek™ has a suite of BLAST tools for online sequence search and retrieval from public databases and proprietary local databases. This tool supports blastn, blastp, blastx, tblastn and tblastx searches. Search results can be viewed in a customized interface. A range of standard search parameters are provided in search options. Customized search parameters can be further added to BLAST suite.



Local BLAST Manager

Local BLAST Manager lets you build and manage databases of sequences and perform powerful BLAST searches from Genchek™ using an ordinary desktop or laptop computer. This tool lets you perform text searches of your databases, as well as offering a wide range of BLAST programs, including blastn, blastp, blastx, tblastn and tblastx. The local BLAST database is highly customizable in order to meet research requirements of your lab.



Six - Frame Analysis and ORF Finding Tool

A researcher can analyze the sequence in all six coding frames (-3 to +3), facilitating a better picture of the query sequence. The ORF finding tool is handy, and can quickly scan for putative open reading frame in a sequence or a collection of sequences.

Gene Finder

This tool can help predict possible genes in a sequence using a neural network algorithm. A neural network system is used to train the tool with annotated data sets. Coding potential for an unknown sequence can be determined thereafter. A researcher can also analyze putative promoter regions, acceptor splice sites, donor splice sites, CpG islands, exons, start and stop codons. A combination of coding potential and other characteristics of a putative region as mentioned above can help determine whether a particular sequence may be part of a gene.

Vector Trimming

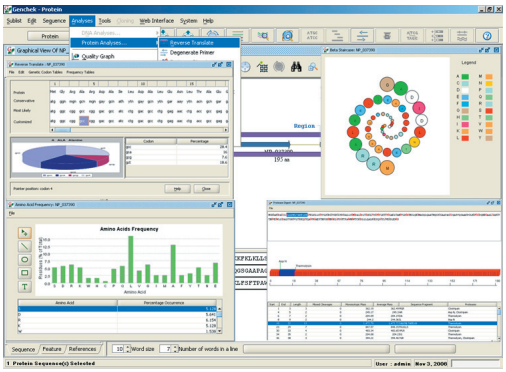
A sequence obtained from a sequencing machine (raw sequence) or from a database may have traces of the vector that was used to clone a fragment of DNA before sequencing. Using this utility, user can remove traces of vector from sample sequence before it is analyzed. Graphical representation of vector helps the researchers to understand the vector and its annotation. This is an intuitively designed viewer where sequence can be viewed in both graphical as well as textual format at the same time.

Protein Analysis

Protease Digest

This tool simulates complete and partial digestion of protein by proteases. User can vary parameters like minimum fragment mass, fragment length, maximum missed cleavage etc. according to the need. The descriptive information associated with

each protease is displayed in a sortable tabular format. On selecting a particular row from the table, it highlights the sequence fragment in the graphical and textual representation of the sequence.



Hydrophobicity Plot

This tool enables the user to generate wide variety of hydrophobicity (hydropathy) plots over the entire length of the protein sequence which helps in determining the flexibility, antigenicity and other related properties. User has the flexibility of changing the sensitivity of the plots by varying the window size.

Charge Vs pH Plot

This utility in Genchek™ plots the titration curve or charge vs. pH curve of the protein. The plot can be used to determine the isoelectric point (pH at which the charge on protein is zero) of the protein. Since different proteins have different pI, by adjusting the pH to pI of a particular protein, one can separate the respective protein in the form of precipitate in the solution.

Helical Wheel

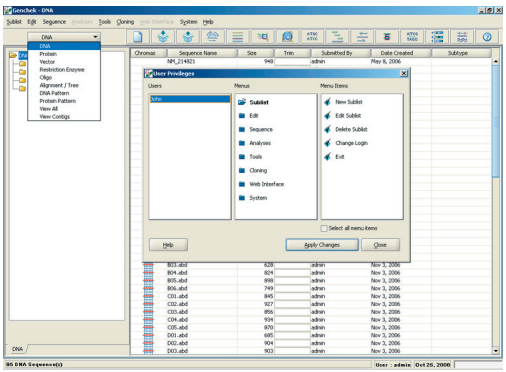
The helical wheel is a way of representing amino acid in a helix, to give an idea where they lie with respect to each other. In the wheel, each amino acid residue is plotted 100 degree apart, as they would exist in 3.6 helix. This utility is used to recognize the amphiphilic character of the specified sequence fragment, which is of particular importance in determining the hydrophobic parts of the sequences.

Generic Protein Statistics

This is one of the most powerful utilities in the protein analysis module of Genchek™. It tabulates the various physico-chemical properties (like percentage of basic and acidic amino acid residues, buried surface area, packing volume, solubility, radius of gyration etc.) of the selected protein.

Data Security and Access Level control in Genchek™

This is an administration utility to control access and access levels in Genchek™. Information about access levels of different users, their usage and privileges is displayed and change of privileges can be done using this tool. Administrator privileges are generally required to exercise access level controls.



Data Management in Genchek™

A robust database architecture with different levels of access serves many purposes such as:

- Centralized data storage, thereby, eliminating the need of multiple data storage disks.
- Data security is taken care of by the database architecture and backup features. A single backup operation can copy and store the entire data to a safe repository rather than taking backup of individual systems and directories therein.
- Interpretation, analysis and report generation is easily facilitated.

Another salient feature of Genchek™ is multiple user environment through client server set up. Genchek™ can be installed on a server and client machines in the network, which can facilitate access to several users simultaneously. In addition to an integral database, Genchek™ facilitates easy export and import of data. Flat files, chromatograms and other standard file formats such as those from GenBank, SWISSPROT, EMBL are supported by and are easily imported to Genchek™. Data are reported internally in XML (extensible Markup Language) format; thereon an application can build a PDF (portable document format), xls (Excel file or spreadsheet) or a text document, thus it can be recognized and read by any application. Similarly, data import from XML format allows other files to be read in Genchek™ .

Why Genchek™ for Your Sequence Analysis Needs?

Genchek™ is designed with user-friendly interface and control, yet the architecture is robust and is compatible with commonly used operating systems/platforms such as Windows 98, Windows 2000, Windows XP, Macintosh, Linux

and UNIX. Genchek™ is very competitively priced. We constantly work with our customers to correctly identify the requirements and integrate customized matrices and algorithms to Genchek™. Thus our customers pay for what they require, instead of packages that may have a few required tools, but are inclusive of costs for those tools that are not required.

The Last Word

Laboratories that are engaged in genomics research have enormous information management requirements that are growing with creation of more knowledge as well as development in technology. Increase in efficiency of information management leads to overall improved efficiency and productivity, thus researchers can focus on discovery per se rather than on mundane archival and tracking tasks. Genchek™ is a comprehensive biological sequence analysis system developed and implemented by Ocimum Biosolutions, India, that meets standard requirements of most biological research laboratories and is customizable, scalable and expandable to meet specific requirements of any individual laboratory.

United States: Ocimum Biosolutions Inc. Fortune Park VI, 8765, Guion Road, Suite G, Indianapolis, IN, 46268, USA
Phone: +1 317 228 0600, Fax: +1 317 228 0700, Email: us@ocimumbio.com

India: Ocimum Biosolutions Ltd, 6th Floor, Reliance Classic, Road No. 1, Banjara Hills, Hyderabad 500 034, A.P., India
Phone: +91 40 666 27200, Fax: +91 40 666 27205, Email: india@ocimumbio.com

The Netherlands: Isogen Biosolutions, (An Ocimum Biosolutions Company), Lagedijk Noord 18, Postbus 220, NL -3400 AE IJsselstein, The Netherlands
Phone: +31 (0) 30 68 78 788, Fax: +31 (0) 30 68 88 009, Email: europe@ocimumbio.com

Ocimum Biosolutions ...enabling R&D™

www.ocimumbio.com