

A background image of a Nepenthes alata plant, showing its characteristic green leaves and red-tipped pitcher traps. The plant is a climbing vine with a long, thin stem and a large, bulbous pitcher trap hanging from the end of a leaf. The background is a soft-focus green, suggesting a natural habitat.

Kodon[®]

Total genome
and sequence analysis
software

**A new light
on the secrets
of life**

Applied Maths

www.applied-maths.com

Kodon

DNA, PROTEIN, AND GENOME SEQUENCE ANALYSIS

In an era of total genome research, Kodon embodies a new generation of sequence analysis software that combines the power of a high-end server with the convenience of a desktop application. Kodon is capable of managing the largest databases, and enables researchers to analyze entire chromosomes and genomes with unparalleled power and speed. With its impressive range of applications, Kodon offers numerous pioneering features, which are designed for use in genome sequencing projects of any size. Some of Kodon's key features are:

- Fully integrated analysis platform, from raw sequencer chromatograms to megabase sequences, with full back-tracking functionality.
- GenBank and EMBL are Kodon's native formats. No conversion, no loss of features and qualifiers from documented sequences.
- Project-oriented multi-user database environment; unlimited capacity. Choice between local, file-based databases and server-based Oracle, PostgreSQL, Microsoft SQL Server, Access, and other database engines. Possibility to connect to existing databases.
- Storage and full analysis capability on local computers. No need to upload data over the Internet, avoiding network traffic, enhancing data security, and ensuring faster return of results.
- Automatic updates of databases from public servers.
- Advanced structured queries of any complexity for searches in local or external databases.
- Rich and uniform sequence annotation display in all analysis tools: vector cloning, ligation, primer design, alignment, homology search, match and repeat analysis, ...
- Feature matrix to search and select related features out of chromosome sequences and analyze them by multiple alignment and clustering.
- Chromosome dot plot matrix to visualize homologous regions in and between chromosomes. Automatic search for repeats and duplicates to locate paralogous and orthologous genes.
- In-silico cloning and restriction analysis

with unparalleled functions and rich, publication-ready output possibilities. Simulation of agarose gels with restriction fragment patterns and size markers.

- Primer design on sequences, chromosomes, and multiple alignments of DNA or protein sequences. Temperature and degeneration plots, search for distinctive primers for PCR detection, and many other features.
- Prediction of RNA and protein secondary structure, helical wheel analysis, motif search, prediction of Leucine zippers, and physico-chemical property analysis.
- The finest multiple alignment editor that exists, offering a wealth of alignment, clustering and phylogenetic inference functions.
- Computationally intensive tasks such as multi-chromosome mapping run in background threads and can be interrupted and resumed whenever desired.
- Script language for automation of tasks, import & export, generation of custom reports, personalization of interface, etc.

Kodon consists of a Basic Software and three modules for specific applications. The main functionality of the modules is as follows:

- **Kodon Basic Software:** contig assembly, database and advanced querying, sequence editor, restriction enzyme analysis, homology search (BLAST)
- **Multiple Alignment, Clustering and Phylogeny:** sequence selection from feature matrix, multiple Alignment, clustering and phylogeny (UPGMA, Neighbor Joining, Parsimony, Maximum Likelihood)
- **Molecular Analysis:** vector construction, primer design, frame analysis, RNA and protein secondary structure, pairwise matching and repeat analysis, motif search and physico-chemical properties analysis
- **Chromosome Mapping:** side-by-side comparison of genomes and chromosomes, analysis of organization and functional behaviour of genomes, chromosome annotation

The full Kodon functionality is physically contained in the same program unit, which guarantees perfect integration of the modules, a consistent user interface, and no tedious switching between programs.

Kodon Basic Software

Kodon is built on a robust database system, supporting the storage and retrieval of tens-of-thousands of sequences of any size. The user has the choice to create local, file-based databases for size-limited projects, and Server-based databases using Oracle, SQL Server, PostgreSQL or other engines, for site-wide access and management of sequence databases of unlimited size.

Queries can be composed based upon database fields, sequence headers, subsequences, etc. The smart subsequence searching tool allows the user to find any subsequence in a database, including a predefined maximum number of gaps and mismatches. Individual query components can be combined into composite queries of

any complexity using logical operators. Queries are nicely represented in a smart interactive diagram and can be saved.

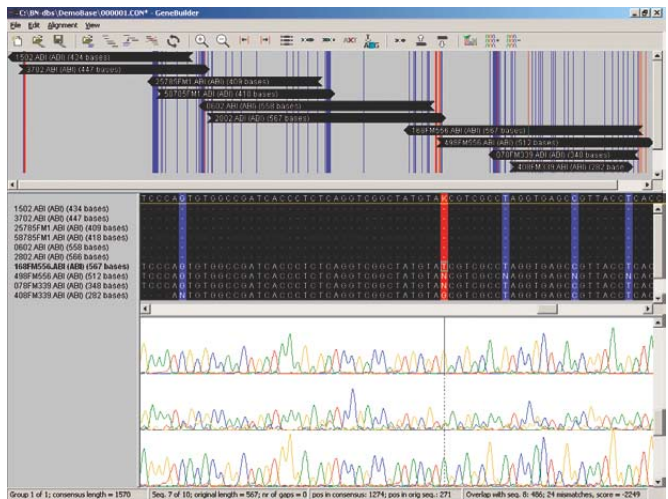
Kodon recognizes GenBank and EMBL as its own native formats. Since no conversion from these major public databases is needed, there is

no loss of any feature or qualifier from fully documented sequences. Multisequence downloads of any size can be instantaneously visualized and analyzed in the database. Other formats such as FastA, GCG are also supported for import and export through script-mediated functions. For quick and easy sequence import from the Web, Kodon has its own, user-friendly web browser. Adding sequences to the database is as easy as clicking a Save button.

Besides the powerful sequence database platform, Kodon's Basic Software offers the following functions:

Contig assembly

The fully featured and easy to use contig assembly editor *GeneBuilder* allows direct import of raw chromatogram files from automated sequencers (e.g. ABI, Beckman, Amersham). To make visual inspection easier, chromatograms of subsequences can be shown simultaneously in aligned mode. Entire projects are displayed with aligned subsequences, aligned chromatograms, and contig overview simultaneously. Further useful features include a multilevel undo function, automatic trimming of subsequences to remove regions of poor resolution, trimming of consensus according to predefined primers, and removal of vectors. *GeneBuilder* can be script-driven so that contig assembly can be fully automated. Critical and potential problems are reported. Assembly projects can be viewed and re-edited at any time.

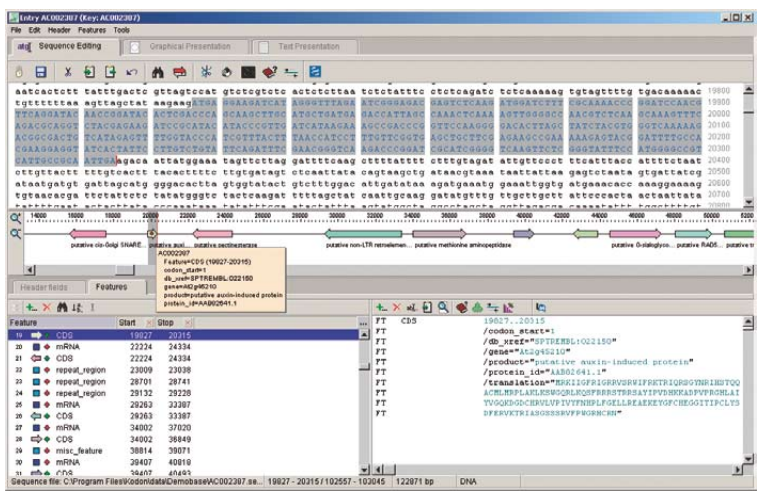


Homology searching

DNA and protein sequences can be screened against Kodon databases as well as downloaded EMBL or GenBank multisequence files. Extremely fast BLAST-based first screening, combined with fine alignment on a restricted result set, results in an excellent tool for screening sequence similarity quickly and with high accuracy. Alternatively, Kodon can launch BLAST queries on NCBI or other Internet databases, and display the results in its own web browser, from which sequences can be saved directly in the database.

Sequence editor

Sequence databases such as EMBL and GenBank can be downloaded directly into Kodon with full preservation of all features and qualifiers in the local database. Sequences can be viewed in three different modes: a standard editing mode combining text, graphical preview and features with qualifiers; a graphics presentation mode (linear or circular)



with annotations of choice, and a text presentation mode showing DNA, translation, and annotations of choice. Features can be added, changed or removed, and new sequences can be annotated. A large number of analysis tools can be launched directly from the sequence or from a selected feature: homology screening, motif search, primer design, restriction enzyme analysis, RNA and protein secondary structure, open reading frame analysis, etc. For selected information in the sequence header or feature descriptions (e.g., citations), instant searches can be launched on PubMed or NCBI.

Restriction enzyme analysis

Advanced restriction enzyme analysis can be performed on sequences up to full chromosome size. The complete REBASE enzyme database is available and can be updated from GenBank or EMBL. Subsets of restriction enzymes available in your lab can be saved. From selected enzyme

Vector construction

The unique vector cloning application is unparalleled in terms of versatility and graphical possibilities. *In silico* experiments can be conducted based upon multiple, stepwise constructed cleavages, fill-ins, ligations etc. The integration of restriction enzyme analysis, and the pos-

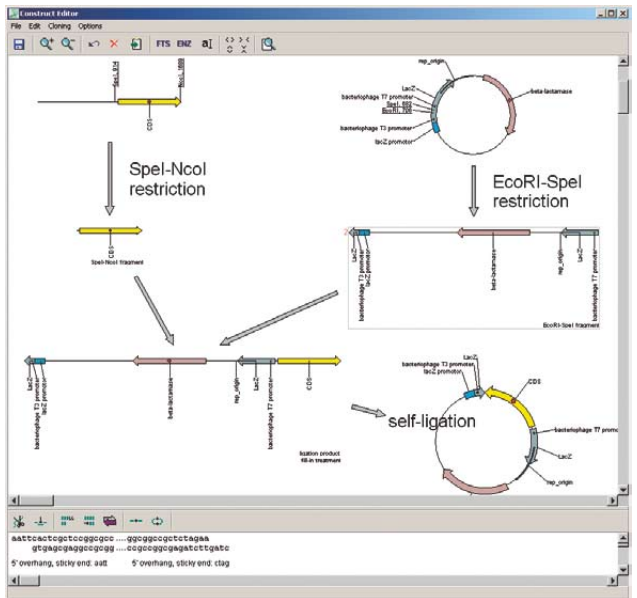
The image displays three overlapping windows from a molecular biology software suite:

- Restriction Enzyme Mapping:** Shows a sequence alignment with various restriction enzyme sites (EcoRI, SpeI, NcoI, PstI, SmaI) marked along a DNA sequence. A table below lists the enzymes selected and their cleavage sites.
- Restriction enzyme identification card:** A detailed card for the XbaI enzyme, including its recognition site (C₆C₆GGC), methylation site (GGCC), and references.
- Restriction enzyme analysis table:** A table showing the results of a digestion experiment, listing enzymes, cleavage sites, and resulting fragment lengths.

Enzymes selected	5' Cleavage	Fragment	3' Cleavage	length
PstI	cut1: 43228/43228	PstI	cut2: 43558/43554	length: 330 bp
SmaI	cut1: 43228/43228	SmaI	cut2: 43558/43554	length: 330 bp
	cut1: 49256/49255	PstI	cut2: 50025/50029	length: 1569 bp
		SmaI	cut2: 52424/52420	length: 1599 bp
		SmaI	cut2: 52426/52420	length: 820 bp
		SmaI	cut2: 54242/54246	length: 1018 bp
		PstI	cut2: 56117/56113	length: 1855 bp
		SmaI	cut2: 59044/59048	length: 1927 bp
		PstI	cut2: 62899/62895	length: 4855 bp
		PstI	cut2: 66221/66217	length: 3222 bp
		PstI	cut2: 71994/71992	length: 5763 bp
		PstI	cut2: 77916/77912	length: 5950 bp
		SmaI	cut2: 77947/77951	length: 51 bp
		SmaI	cut2: 89200/89204	length: 11253 bp

Molecular Analysis

Kodon's Molecular Analysis module is the perfect supplement to the Basic Software for molecular genetics research. Besides its remarkable vector cloning application, the module comprises complete solutions for primer design, frame analysis, repeat and match analysis, motif search, RNA folding, protein secondary structure, and physico-chemical properties analysis.



sibility of obtaining predicted gel patterns for any step of the experiment, as described above, enables the user to design and monitor experiments with unprecedented efficiency. Obtained constructs can be used to conduct new ligations within the same project.

Any intermediate or end-product in a vector cloning project can be analyzed as a sequence entry and saved in the database. Since sequences in a vector cloning project are descended from parent sequences in the database, any changes made to these sequences, and their consequences, are effectuated in the project.

The entire experiment is displayed as a professional, ready-to-publish graphical presentation. All graphical layout and annotations defined for the parent sequences are preserved in the vector cloning project.

Primer design

Design of primers is an essential part of molecular genetics research, and therefore, this application has been equipped with a large number of features. User-adjustable settings allow primers and PCR products to be found that match any combination of conditions. These settings include the preferred primer size, preferred melting temperature, maximum degeneracy, salt and DNA concentration, all of which can be weighted for relative importance. Settings can also be adjusted and search results updated in real-time using temperature and degeneration plots. The software screens for potential hairpin structures and

dimers. Searches can be done at DNA and protein level, on sequences of any size. Searches can be narrowed to find primers that amplify a specific region, for example a CDS. The inclusion of specific restriction enzyme sites can be requested.

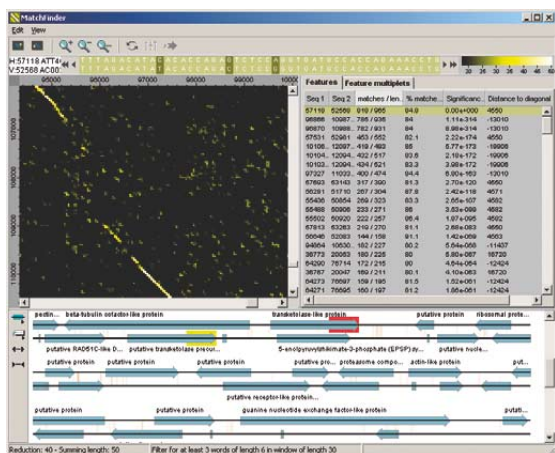
Primer searches can also be performed on multiple alignments. Kodon will screen for primers and PCR products that are compatible with all the sequences that are part of the alignment. A unique feature of Kodon is its possibility to search for discriminative primers, PCR products, or probes between sets of sequences. This tool has proven to be of great value for the design of primers or probes that detect specific target organisms, for example human, animal or plant pathogenic bacteria or viruses.

› Frame analysis

Kodon supports all translation tables currently published. Updated tables can be downloaded directly from GenBank or EMBL. Frame determination is made easy and convenient with a six-frame graphical presentation on the annotated sequence. The plausibility of each hypothetical protein encoding sequence is indicated, based upon codon usage frequency tables available for most classes of organisms. As more codon usage tables become available, they can be added to the system.

› Repeat and match search

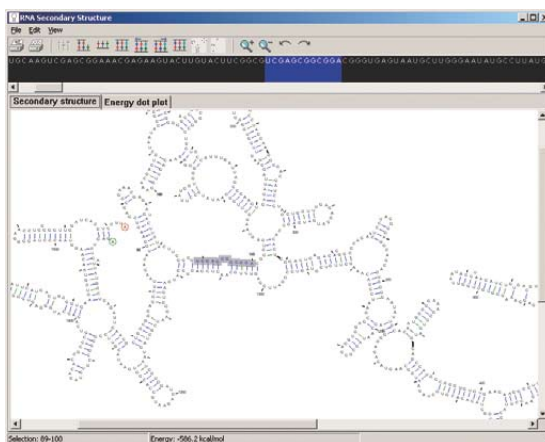
An advanced pattern matching engine will find all repeats and inverted repeats within a sequence of any size, including entire chromosomes of millions of bases. Likewise, the engine will find all matches between pairs of sequences or chromosomes as well. Matching regions are graphically represented in a dot plot matrix, and are listed in a table. They can be sorted upon significance, length, percentage of matching bases etc. The software will also find multiplets of matches, e.g., indicating repeat regions. Additionally, Kodon contains a specialized searching tool for short exact repeats. This tool is of great value for the de-



tection of new tandem repeats, interspersed repeated DNA, transposons, microsatellites, insertion sequences, etc.

› RNA secondary structure prediction

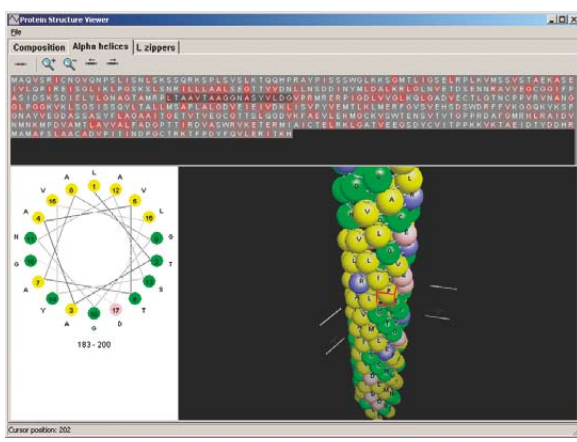
RNA secondary structure is predicted using a nearest neighbor thermodynamic folding model. Various parameters such as temperature, tetra loops, dangling end energies, non-standard base pairing, etc. can be specified. Selections on the sequence made by the user can be forced to be unpaired or paired upstream, downstream or with a specific target selection. The obtained secondary



structure model can be scaled, rotated, moved with the mouse, and printed or copied as enhanced metafile. In addition, an energy dot plot can be shown and printed, indicating how the algorithm has obtained the proposed secondary structure of the sequence.

› Protein structure and property analysis

Protein secondary structure. Alpha helices can be predicted on peptide sequences using the *helical wheel* presentation. In addition, advanced three-dimensional alpha helix models can be visualized and printed.



Kodon analyzes and displays potential Leucine zippers, relaxed Leucine regions, and coiled coils or supercoils, using different colors.

Compositional analysis. Amino acids can be divided into classes such as hydrophobic, hydrophilic, neutral, aromatic, acidic, and basic, as well as user-defined classes, of which frequency plots can be composed.

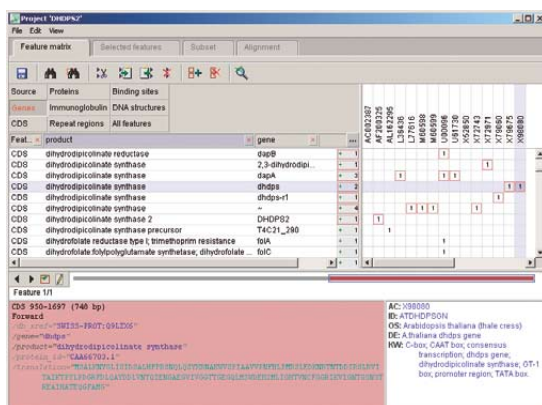
Motif search. Kodon will search for all known motifs (conserved patterns) in protein sequences, display the characteristics and description for each selected motif, and display information about the type of protein site. Through its compatibility with Prosite™, the motif database can be kept up-to-date with public databases with a simple download.

Multiple Alignment and Cluster Analysis

The capacity to align many thousands of DNA or protein sequences, the intuitive graphical user interface, and the large number of unique features and functions, together form the key characteristics of Kodon's powerful sequence alignment and clustering module. The multiple alignment editor interacts seamlessly with dependent applications such as alignment-based primer design.

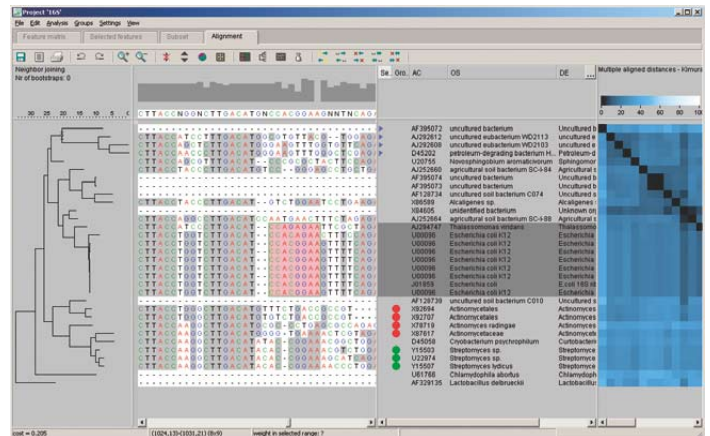
Feature matrix

Multiple alignment projects can be generated from queries or manual selections, which can be saved and stored with all settings, editing, and calculations done. Along with a multiple alignment project, a surveyable table allows the user to select any feature or combination of features out of chromosomes or sequences, to be used for further analysis. For example, one can select all types of kinase enzymes from a list of chromosome sequences and perform a multiple alignment to locate conserved regions in this type of enzymes. In case a gene occurs more than once in the same organism, one or more copies can be selected to be included in the multiple alignment.



Multiple alignment

A very powerful sequence alignment engine makes it possible to obtain alignments of up to ten thousand sequences, each many thousands of bases long. Optimization parameters allow the alignment to be tuned to maximal accuracy and speed. The alignment of protein sequences relies on an editable scoring table. For protein encoding DNA sequences of low relatedness, the software has a function to align the translated protein sequences, and superimpose the obtained protein alignment on the DNA sequences. This approach ensures the most reliable alignment of coding sequences, yet preserving the sensitivity of comparing DNA sequences.



The multiple alignment editor offers a number of unique edit functions, such as the possibility to select blocks spanning a region on a range of sequences, easy manual drag-and-drop realignment of any selected region, a multi-step undo and redo function, etc.

Ingenious tools are available for partial controlled alignment, e.g. to update alignments within selected regions/sequences, or to realign a region using other cost settings. With the *incremental alignment* algorithm, new sequences can be added to an existing multiple alignment at any time in a controlled way, guaranteeing that any manual correction work done previously is preserved.

A consensus sequence can be calculated from the entire alignment or from any selection, according to user-defined threshold and confidence settings. A weight factor can be set for each position on the consensus. Using this feature, the user can assign zero weights to regions to be excluded. Another useful feature is the automatic weight assignment, whereby the software calculates the entropy for each consensus position based upon the multiple alignment; the lower the entropy for a position, the higher its weight. Conserved positions in the multiple alignment are thus assigned a higher weight than variable positions. When used in incremental alignments, conserved positions will have a higher impact on the alignment of additional sequences, which allows for more reliable alignment of additional

sequences to an existing multiple alignment.

A search tool allows subsequences to be found throughout the multiple alignment. The user can specify a maximum allowed number of mismatches, or can enter degenerated positions according to the IUPAC convention.

Cluster analysis and phylogeny

Available clustering methods in Kodon include Unweighted Pair Group clustering, Neighbor-Joining, Generalized Parsimony, and Maximum Likelihood. Maximum parsimony solutions are calculated using a superior method called *Simulated Annealing*. Confidence indication of dendrogram branches is obtained using bootstrap analysis. For clusterings with a phylogenetic dimension, two implementations for evolutionary correction can be applied to the distance scale: the *Jukes & Cantor* and the *Kimura 2 parameter* method.

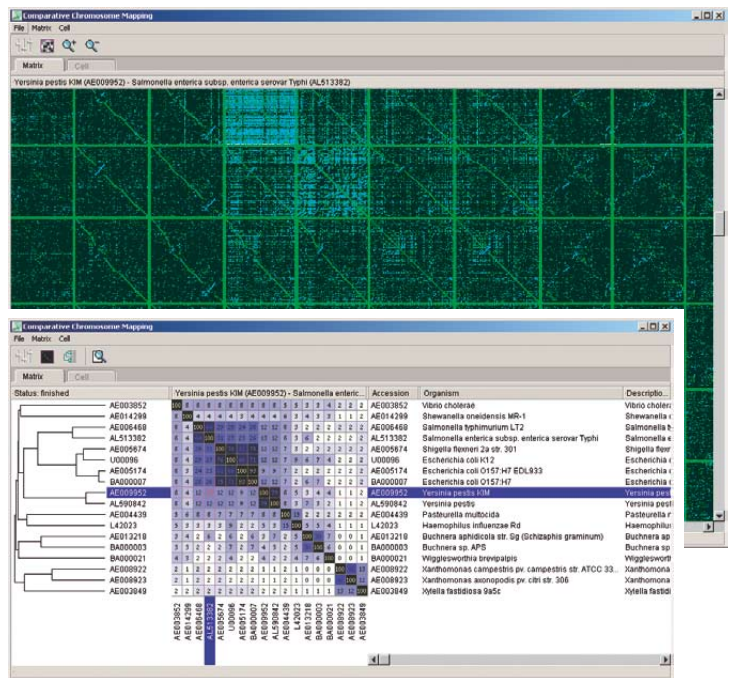
A large number of editing tools are available to facilitate the interpretation of large alignments and clusterings: individual sequences or groups of sequences can be moved up or down, branches of dendrograms can be swapped, unrooted trees can be rerooted, groups of sequences can be selected directly from dendrogram branches, groups of entries can be created and assigned different color codes, and a reference sequence can be defined for consistent base or amino acid numbering.

Any two sequences from a multiple alignment can be subjected to a detailed pairwise comparison, launching Kodon's Repeat and Match search application. The multiple alignment editor also forms the basis for primer design on aligned sequences: consensus primers can be found for any selection of aligned protein or DNA sequences, and discriminative primers can be searched for between any two sets of aligned sequences.

Kodon's multiple alignment editor offers powerful graphical output possibilities. A click on a button turns the editor into a print preview window, showing the layout of the pages to print. Every individual component can be added to or deleted from the graphic, or its size can be changed. Graphics can be sent directly to the printer or copied to other applications as enhanced metafiles. As an interesting alternative for publication purposes, alignments can be saved as a Rich Text Format (RTF) files, which allows the alignment to be placed and formatted in a document.

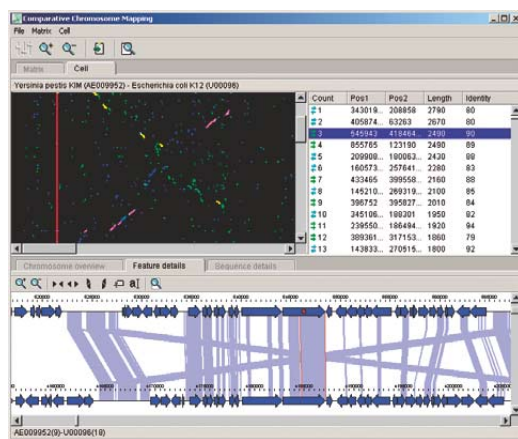
Chromosome Mapping

The comparative chromosome mapping module is designed to perform in-depth side-by-side comparison of full chromosomes. The uniqueness of Kodon's chromosome mapping lies in the fact that it can compare any number of



chromosomes with one another, resulting in a matrix of $n \times n$ dot plot matrices. This invaluable source of information provides a better insight in the organization and functional behaviour of genomes than ever possible before. Based upon the percentage of matches between chromosomes, the software is able to construct similarity matrices and dendrograms, depicting the global relatedness between organisms based upon their full genomes.

The possibility to match any number of chromosomes or sequences, annotated or not, with one another, provides a sophisticated way of annotating sequences and yields information that could never be resolved before, such as the chromosomal order of genes, unique or ubiquitous genes in organisms, metabolic pathways, homologous genes, etc.





Applied Maths

Keistraat 120, B-9830 Sint - Martens - Latem, Belgium
Phone +32 9 2222100 • Fax +32 9 2222102

512 East 11th Street, Suite 207, Austin, Texas 78701, USA
Phone +(1) 512-482-9700 • Fax +(1) 512-482-9708

www.applied-maths.com

Kodon is a trademark of Applied Maths BVBA.
All other trademarks are the properties of their respective owners.
The information in this brochure is subject to changes without prior notice.
Copyright 2003, Applied Maths BVBA.
All rights reserved.