BioNumerics 7.5 consists of:

- **5 data modules**: *Fingerprint Data*, *Character Data*, *Sequence Data*, *Whole Genome Map Data,* and *Trend Data.*
- **5 analysis modules**: *Tree and Network Inference, Genome Analysis Tools, Classifiers and Identification, Dimensioning techniques and Matrix Mining,* and *Audit Trails*

Each analysis module can be combined with any or all application modules, so you can benefit from your most optimal configuration. The descriptions below list most, but not all functions and possibilities of the BioNumerics modules. We provide both standalone as network licenses for each configuration. Please contact us for details and prices.

Data Modules

General functionality

Database. Object-oriented relational multi-user database interface. Store information as text, number or date. Lock information fields or limit editing options by using pick lists. Easy drag & drop linkage of multiple experiments to database entries. Powerful search engine for combined database searches on information fields and experiment presence and/or contents (character values, sequences, ranges, bands, etc.). Creation of object queries to retrieve data from any database object. Storage and management of database queries and external attachments. Multi-database system: each database can contain any combination of different experiment types. With an optional leveled database setup, a richer hierarchical data structure is created to reflect the real life situation better and to deal efficiently with replicate experiments. Creation of XML files from any selection of entries and techniques in the database for export. XML files can be imported as fully editable database entries. XML exchange is the preferred way of exchanging database entries in a peer-to-peer network.

Interface. Customizable GUI with dockable panels and editable toolbars. Multi-window interface allows multiple instances to be opened from the same and different applications. Pop-up navigator to navigate between BioNumerics applications.

Chart Tools. Powerful and versatile charting tool, allowing charts, graphs and tables to be generated for all sorts of text based and numerical information from the database. Chart types include profile charts, profile difference charts, bar graphs, histograms, scatter plots, frequency bar graphs, box & whiskers charts, contingency tables, ANOVA charts, component summaries, pie charts, etc. Rich presentation, printing and exporting tools. In combination with the *Dimensioning techniques and Matrix Mining* module, appropriate statistical tests can be performed on each chart type.

ODBC Connectivity. Possibility to link BioNumerics to ODBC compatible high performance database management systems such as Oracle, SQL Server, MySQL.

User Management. Comprehensive set of User and Security tools, including creation of Users with logins and passwords and User Groups defining specific privileges. Control password timeout and strength, user activity logging and data input consistency. Possibility to define access privileges for each individual database object. Create, Modify, Delete, Sign, Restore, Lock and Unlock privileges can be granted to specific users.

Geographical Coding. Perform geocoding and plot database entries on an interactive geographical map based on locations present in the database.

Fingerprint Data (FPR)

Gel Image Processing and Normalization. Input of any bitmap images, densitograms, and chromatograms of unlimited file size. Image pre-editing and cleaning. 3D representation of bitmaps. Automatic lane finding, gel strip borders and tracking splines adjustable for individual lanes. Automated and manual alignment by pattern recognition using external reference patterns and/or internal reference bands. On-screen normalization of bitmap images with indication of reliability and possible misalignments. Adjustable background subtraction and curve smoothing. Spot removal. Direct comparison of patterns normalized with different reference systems. Band-search algorithms with adjustable sensitivity for shoulder and double-band finding. Quantification of molecular sizes or any other metric unit using linear, logarithmic, combined logarithmic-third power regression, cubic spline or pole functions. Accurate expression of protein or nucleic acid quantities or concentrations based on cubic spline regression using known calibration peaks.

Processing of Genetic Analyzers' Data: Input of electropherograms, generated by genetic analyzers and other multi-channel capillary electrophoresis equipment. Visualization of individual and combined channels as horizontal curves or pseudo-gel lanes. Intuitive zooming and navigation. Flexible peak detection algorithm, with optional filters for stutter peaks, spikes, noise and peak doublets. Visualization of bleed-through areas (pull-up peaks) and corrected peak detection. Dedicated algorithm for normalization using reference dyes. Commercial size markers can be picked from a list. Distortion bars to assess normalization quality.

Band Matching Analysis. Display of any combination of normalized 2D-bitmap strips, densitograms or reconstructed patterns in comparison window. Search for discriminative bands/peaks between selected groups of patterns; search for unique and common bands/peaks. Possibility to mark bands/peaks as uncertain. Binary and quantitative band matching tables of multiple combined fingerprints. Possibility to define named band classes based upon size and position (e.g. for DGGE/TGGE analysis). Add/edit bands directly in the comparison window in band matching mode.

Spectral Data Processing, Filtering, Summarizing and Peak Detection: Import of spectrum data from different formats (mass-intensity list as txt, peak table, btmsp and mzML). Preprocessing of raw spectrum data: resampling, background subtraction, noise calculation and filtering, trimming, peak detection and peak filtering. Default templates for preprocessing are available that have been optimized for spectra from whole bacterial protein extracts (adaption of Continuous Wavelet Transformation algorithm). Flexible visualization of both raw and processed data and summary data. View spectra in overlay, under each other, visualize different preprocessing steps and many more. Generate and export peak lists from raw spectra. Perform a peak matching for further analysis: finding discriminative peaks, common peaks, identify unknown samples based on spectra,…

Character Data (CHR)

Flexible import tools and programmable routines for the import of character data from text files and ODBC-compatible sources (MS Excel®, MS Access®,…). Character data may include any existing data set, numerical, binary or continuous, within any range, with fixed or variable number of characters. Character names may be given by the user or automatically imported. Unlimited length of character arrays. Possibility to map character values to categories according to predefined criteria. Direct digitization and processing of micro-arrays, test panels, microtiter plates, dot blots, etc. from TIFF files. Character profiles can be displayed in a panel with user-defined representations and color scales or in a list with values, mappings, colors, or combinations thereof. Display of truthful image of any test panel and easy on-screen data input. Import of similarity or distance matrices. Partial matrices accepted (e.g. DNA homology matrices). Unlimited matrix size.

Sequence Data (SEQ)

Sequences. Direct import of sequences from EMBL, GenBank, Flat A, FASTA formats and high-throughput sequencing data. Import of nucleic acid and amino acid sequences. Easy paste from clipboard, and manual editing. Project-based contig assembly and consensus editing from sequencer chromatogram files (ABI, Beckman, MegaBace). Full IUPAC code support for consensus naming. Contig projects can be opened from entry editor, comparison and multiple alignments. SNP detection and analysis (in combination with the Tree and Network Inference module). Powerful automated batch-assembly for high-throughput sequence processing. Import of sequence assemblies in BAM and SAM format. Circular genome viewer for interactive exploration of sequence data.

Sequence Read Sets. Import of any base space encoded sequence data from 454 GS systems (Roche Applied Science). FASTA files, FASTQ files from e.g. the Genome Analyzer, MiSeq or HiSeq (Illumina), the SOLiD systems (Applied Biosystems), the Ion Torrent PGM (Life Technologies), the PacBio RS (Pacific Biosciences) and the HeliScope system (Helicos BioSciences). Imported sequence read sets are linked to new or existing database entries. Based on predefined charts providing insight in the sequence quality and the possible presence of sequencing artifacts in the data, assessments can be made for the required preprocessing steps (demultiplexing, splitting 454 paired-end reads, trimming, chimera detections, primer removal, sequence selection) before starting the actual analysis (whole genome *de novo* assembly or resequencing assembly of bacterial genomes, and deep sequencing metagenomics analysis of a phylogenetic marker (e.g. 16S rRNA) and single sample diversity analysis).

Power Assembler. Allows creation of project-based pipelines for resequencing analyses and *de novo* sequencing to assemble contigs up to full genome size from millions of reads. From simple and easy to advanced and customized workflows, through predefined or user-defined analysis pipelines. Graphic support to trim and filter the reads using different criteria. Combining single and paired end reads, or reads from multiple technologies into one project. Data analysis using multiplex identifiers. Detailed analysis report with summary graphs and statistics. Creates graphical sequence curve analyses. Exports results to the BioNumerics database. Power Assembly project can be opened from the dedicated analysis windows, with selected base highlighted (sequence viewer, comparison, alignment, chromosome comparison, annotation). Different levels of interaction with the user to define e.g. parameter settings.

BLAST Analysis. Allows to compare amino-acid sequences of proteins or nucleotides of DNA sequences. Different BLAST algorithms are available to perform different sequence comparisons, e.g. a DNA query to a DNA database, a protein query to a protein database, and a DNA query, translated in all six reading frames, to a protein sequence database. Other adaptations of BLAST, such as PSI-BLAST and RPS-BLAST perform comparisons against sequence profiles. Local BioNumerics BLAST databases as well as the online available NCBI BLAST repositories can be used as search set. The resulting BLAST sequence hits can be imported in the BioNumerics database.

Frame Analysis. Find all open reading frames (ORF) and predict protein coding sequences (PCS) on a sequence.

Restriction Enzyme Analysis. In-silico multi-purpose analysis of restriction enzyme cleavage suitable for cloning experiments as well as for RFLP, PFGE and AFLP design.

Primer Design. Search for optimal primers or primer combinations for the most diverse experiment setups by taking into account various experimental parameters. Includes differential primer search to design primers or probes that discriminate selected sets of sequences from each other.

Whole Genome Map Data (WGM)

Analysis of high resolution ordered whole genome restriction maps, mainly focused on strain typing and characterization. Straightforward XML-based import of whole genome maps from the Argus™ Optical Mapping System from OpGen® (www.opgen.com).

Comprehensive Viewing & Search tools. Direct and inverted match indication of fragments for fast discovery of genomic differences (indels, duplications) across a series of whole genome maps. Zoom function into a specified range for detailed viewing at the individual fragment level. Fragments can be annotated with a customizable label for further reference. An intuitive search function picks up whole genome map fragments according to their size, label, etc. A convenient highlighting tool allows a quick, at-a-glance discovery of fragments resulting from a search action.

Map-Based Clustering. Distinguishing highly related strains through fast and accurate map-based clustering using new algorithms based on size tolerance or pattern based matching of whole genome map for calculating similarities (only in combination with the *Tree and Network Inference* module).

Genomic Differences at-a-glance. Pairwise and multiple alignment of whole genome maps, in combination with the highlighting, annotation and alignment options, allow to accentuate genomic differences like inversions or deletions (only in combination with the *Tree and Network Inference* module).

Finding Discriminating Fragments. Exclusive tool for finding discriminating fragments on whole genome maps for a (group of) isolate(s) will make an inventory of such fragments, which can be readily flashed-out by the highlighting tool. Fragments discriminating between two or more groups of isolates is made possible through so-called *pattern match classes* and have the ultimate goal of biomarker screening for a specific group of isolates. Further in depth analysis such as PCA or matrix mining in combination with the *Dimensioning techniques and Matrix Mining* module.

▪ *Trend Data (TRD)*

Analyzes series of readings in function of a changing factor (time, temperature, etc.), which define a trend. Examples are bacterial growth curves, kinetics of metabolic and enzymatic activity measurements, real-time PCR, or time-course experiments using microarrays. Import from text files and Omnilog csv files. Mathematical fitting using any of twelve different models, including Logistic growth, Gompertz, Gaussian, Hyperbolic, Power, Exponential, etc. with automatic parameter calculation, useful for analysis and comparison. User can add custom parameters such as statistic parameters, slopes, and values at fixed X. Comparison and clustering can be done on a selected parameter or a combination of multiple parameters. Comprehensive plotting tools with color and group indications.

Trend Regression Analysis. In combination with the *Dimensioning techniques and Matrix Mining* module, advanced *Trend regression analysis* can be performed to detect correlations and cross-correlations between multiple trend-based variables.

Analysis Modules

▪ *Tree and Network Inference (TNI)*

Methods. Comparisons of up to 20,000 database entries, various similarity/distance coefficients for different data types: Pearson product-moment correlation, cosine correlation, Dice or Nei and Li, Jaccard, Jeffrey's X, Ochiai,... Fuzzy logic and area sensitivity for banding patterns. Adjustable trace-to-trace optimization and tolerance settings for banding patterns. Statistical determination of most justified tolerance settings for banding patterns. Gower, Canberra metric, Simple Matching, etc. for character data. Categorical coefficient for multi-state data (VNTR, MLST, AB resistance patterns, etc.). Interactive wizard-driven input of parameters, options and choices. Predefined and user-defined cluster analysis templates in the advanced clustering window make cluster analysis more intuitive for users with little statistical background.

Similarity-Based Clustering. Unweighted Pair-Grouping (UPGMA), Weighted Pair-Grouping (WPGMA), Complete Linkage (furthest neighbor), Single Linkage (nearest neighbor), Ward, Centroid and Median clustering, Correlation Eliminator and Partial Correlation Eliminator.

Phylogenetic Inference Methods. Generalized Parsimony, Maximum Likelihood, Neighbor Joining, Bio-Neighbor Joining, NeighborNet. Population modeling: Analysis of categorical data such as MLST or VNTR (MLVA) using Minimum Spanning Trees to reconstruct evolutionary models. Advanced presentation and editing tools, including faithful tree representation ('rendered trees').

Interpretation. Combined display of character images, sequences, normalized pattern images, with similarity matrices and sorted according to dendrogram(s). Indication of statistical error at all linkage levels and calculation of co-phenetic correlation. "Seaweed" and pseudo-rooted representation for unrooted trees. Bootstrap analysis for single or composite datasets. Display of sorted similarity matrices, shaded or with numerical similarity values. Impressive edit and publishing functions. Enhanced presentation and printing facilities, in a WYSIWYG environment. Direct interaction between database and dendrogram. Incremental and decremental clustering: new entries can be added to or deleted from existing cluster analyses, without having to recalculate the complete analysis. All features of a comparison can be stored to disk.

Cluster Reliability. Patented method allowing the reliability of clusters to be calculated and visualized for any clustering algorithm and data set. The method enumerates the reliability of dendrograms or networks in function of degeneracies as well as poorly resolved clusters and can calculate consensus trees or networks that impose a minimum reliability threshold on each resolved cluster.

Congruence Between Techniques. Calculation of global similarity or congruence between different techniques as matrix or dendrogram. Easy visualization of taxonomic depth or level of each technique by pairwise regression plots of similarities.

Composite Cluster Analysis. Different data sets of the same type and of different types (fingerprint, character, sequence and matrix) can be combined into one consensus clustering. Calculation of global similarity by merging characters or by averaging experiment-related similarities. Optional weighting based on number of characters or defined by the user.

▪ *Genome Analysis Tools (GAT)*

Multi-Chromosome Alignment. Calculation of NxN matrices of pairwise comparisons among sets of N whole chromosome sequences. Alignments can be both DNA and coding sequence based. Creation of multiple alignments from N full chromosomes, using one chromosome as reference. Various views make comparisons between genomes straightforward and allow them to be studied thoroughly up to base or amino acid level. Visual interpretation of genome rearrangements, syntenies and parallelisms is facilitated by the dot plot matrix, the pairwise alignment view and the detailed alignment view, which are all linked and automatically updated. Starting from the multiple chromosome alignment, the following analyses can be performed: mutation analysis, subsequence searches, feature searches and dNdS calculations.

Annotation. ORF-based annotation of unannotated or partially annotated chromosome by alignment and comparison with one or multiple annotated chromosomes (from related organisms). Display detailed alignment view of query sequence and reference sequence for any selected hit. Annotation projects can be saved for later re-editing. Additional reference sequences can be added to improve the annotation. Full annotation information can be saved in the original sequence or a copy.

Metagenomics. Starting from sequence read data, typical preprocessing can be performed such as trimming and chimera removal in preparation for the actual analysis including clustering of the sequences up to the final visualization of the operational taxonomic units (OTU) abundances and rarefaction curves, or to the evaluation of the alpha diversity using a plethora of indices. The metagenomics functionality includes a fully interactive reporting service for the interpretation and manipulation of the results. The metagenomics functionality developed within the logic of the existing BioNumerics functionality, allows the integration of metagenomics analysis with existing methods already analyzed in this environment. In this way scientists are able to combine the formerly commonly used environmental analyses such as DGGE, TGGE, or ARISA with the newly obtained metagenomics results and to compare these different methods. Additionally, the elaborated functionality present in the Comparison window allows a more in depth comparative analysis of different metagenomics samples.

Classifiers and Identification (IDN)

Database Screening. Fast identification of batches of entries with entire databases or selections from databases, using all available coefficients.

Classifiers. State-of-the-art trained classifiers such as Support Vector Machines and Naive Bayesian Classifiers are used for quick and accurate identification of complex groupings. Includes a cross-validation framework and convenient tools for automatic parameter optimization.

Identification Projects. Creation of highly characteristic identification projects based on existing comparisons. Identification is done using stored classifiers, which can be based on specific similarity measures or trained classifiers. Comprehensive identification reports showing results for each available classifier. Many different viewing options and statistical tools to facilitate interpretation.

Decision Networks. Allows automated workflows to be built that make decisions, predict features, perform queries, fill in fields, create graphs and plots, etc. Can be used for resistance prediction, in breeding research, for complex reporting or for automated analysis of multilevel and polyphasic data analysis and data sorting. Includes the option to build decision trees.

Dimensioning techniques and Matrix Mining (DMM)

Principal Component Analysis. Non-hierarchic grouping by PCA. Spatial representation of clouds of entries in user-definable X-Y-Z coordinate systems. Indication of total discrimination of axes. Real-time rotation of coordinate system to enhance perception of 3-D structures. Advanced Open-GL presentation and layout for publication. Delineation of populations using colors and/or codes. Plotting of dendrogram branches on PCA for advanced grouping, comparisons and methodological validations.

Multi-Dimensioning Scaling. Non-hierarchic grouping by MDS. Iterative optimization of distances according to similarity matrix. Same presentation features as PCA.

Self-Organizing Maps. Non-hierarchic grouping by the technique of Self-Organizing Maps (Kohonen maps), a variant of the neural network approach, extremely useful for large and complex data sets. Includes quick classification tool by placing unknown entries in existing SOM.

(M)ANOVA. Advanced statistical analysis of discriminative features between selected groups with indication of confidence and based on multivariate analysis of variance. Full validity testing of the model including multivariate normality test, univariate normality tests and homoscedasticity tests. Select useful characters directly from the Analysis of Variance report window and mark these characters as active/inactive for analysis and identification purposes.

Statistics. A number of parametric and non-parametric statistical tests can be performed in an easy and intuitive environment (Chi-square test, T-test, Wilcoxon signed-ranks test, Kruskal-Wallis test, ANOVA, Pearson correlation test, Spearman rank-order test). Automatic display of available tests for each input data type. Kolgomorov-Smirnov test for normality.

Partition Mapping. Analyzes the correspondences between two partitions (classifications) and produces a number of mapping rules that define the significantly pairing groups between the two sets. The partition mapping tool is very useful for analyzing the congruences and discrepancies between typing and classification techniques and for defining reliable and consistent groups on the basis of multiple classification methods.

Matrix Mining. Offering a wide range of data analysis tools that can be applied to large character datasets (characters, band classes, peak classes, aligned sequences, ...) including hierarchical clustering, partitioning, principal components analysis, and self-organizing maps. Extensive set of mathematical operations for data preprocessing and normalization (log transformation, averaging values, quantile normalization, ...). Creation of subsets, stepwise reducing the dataset to characters that have an interesting behavior, leaving out those that are invariant or that have unacceptable errors associated with them. Full error handling through all analysis, mining and statistics functions. A variety of statistical tests is implemented and presented in a wizard to guide the non-expert user. Plots can be derived from virtually any combination of data/results in the analysis. The plots are automatically updated when the data is manipulated.

Audit Trails (ADT)

Audit Trails. This module enables all changes to any database object in BioNumerics to be recorded and versioned in audit trails. Trailing strength ranges from logging only to full tracking. Optionally, deleted objects can be prevented from being overwritten. Along with each change made to an object, the previous version of the object is automatically stored including time and user data. The audit trail history can be viewed for any particular object, for specific classes of objects or for all audit-trailed objects in the database. Previous versions of an object can be restored.

Digital Signing. Secure digital signature key pairs can be used by authorized users to sign and validate final processed data entries and/or any further changes made. Checks for the validity of digital signatures and for any fraudulent changes made to the data after digitally signing.

This module allows the entire BioNumerics platform to be used in compliance with the strictest Good Manufacturing Practice regulations, including FDA 21 CFR Part 11.

BioNumerics Licensing Solutions

Standalone Licenses. Single user standalone licenses come with a USB hardware protection key (dongle). The software itself can be installed on multiple computers and the license is activated by attaching the dongle.

Powerful Network Solutions. HTTP-based licensing protocol, ensuring compatibility with most firewall and LAN restrictions. Full information and control of license usage at any time by any user and optional logging of critical events or all actions. Compatible with Windows XP, Windows Vista, Windows 7, Windows Server 2003, Windows Server 2008. Dongle-less (soft-lock) solutions available for network licenses, facilitating the interaction with virtualized servers.

License Limits. Network versions are available from 1 up to any number of users. Contact us for details and pricing.