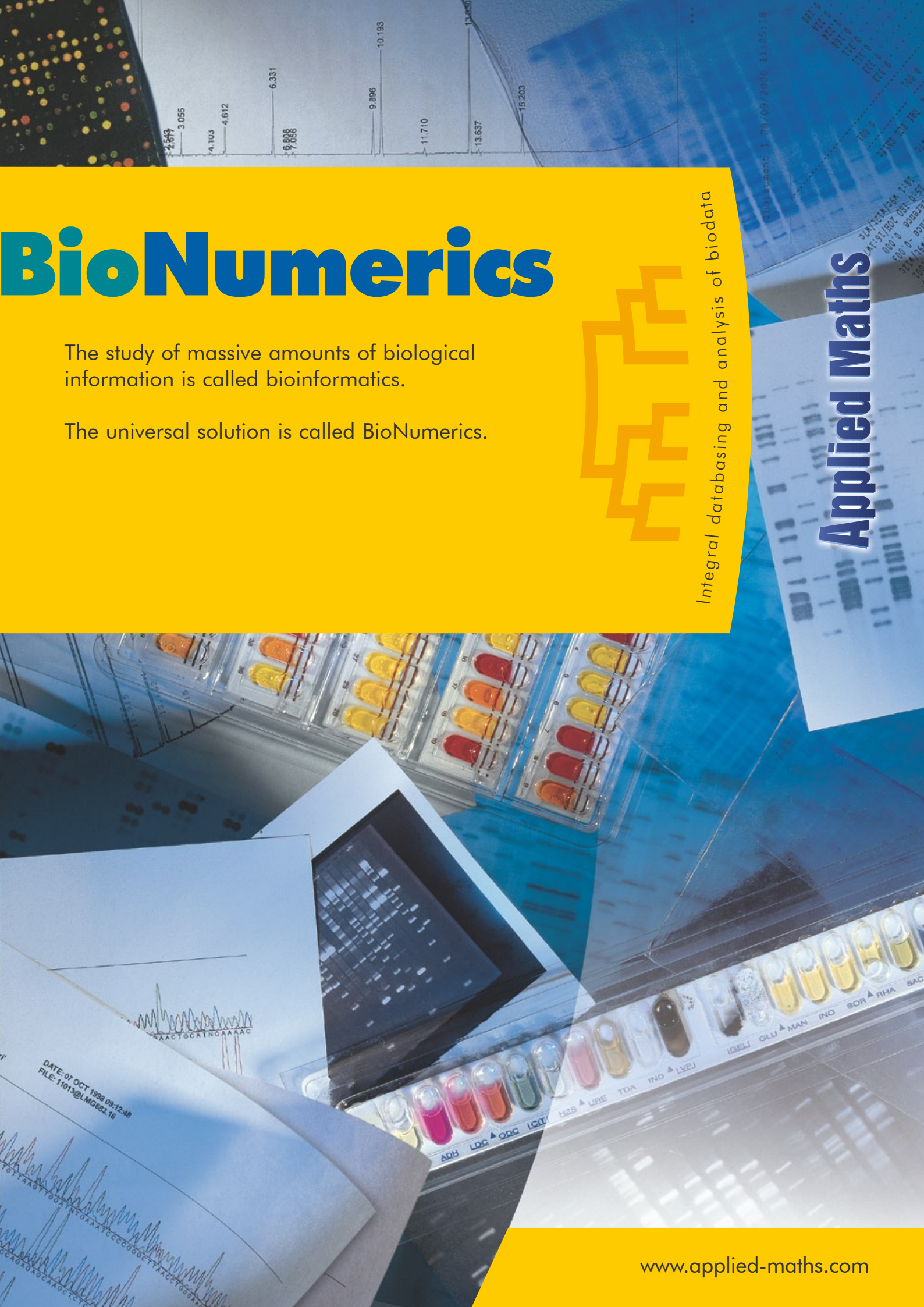

BioNumerics
The study of massive amounts of biological information is called bioinformatics.
The universal solution is called BioNumerics.
Integral databasing and analysis of biodata
Applied Maths
www.applied-maths.com

Organisms, samples

Animals, plants, microbial strains or communities, fungi, tissue, samples, etc.

1-D Fingerprints

Electrophoresis gels, densitometric records, HPLC, spectrophotometry, etc.

Character sets

Phenotypic test panels, antibiotic resistance profiles, microarrays, etc.

Sequences

DNA, RNA, and protein sequences

2-D gels

Two-dimensional protein gels

Input

Import, conversion, image acquisition, normalization of gels, assembly of sequences, etc.

BioNumerics Database

Different experiments and descriptive information linked to database entries

Identification

- Quick similarity search
- Construction of libraries
- Neural networks

Clustering

- Dendograms from individual experiments and composite data sets
- Phylogenetic tree construction

Dimensioning

- Principal components analysis
- Discriminant analysis
- Self-organising maps

Statistical tools

- Cluster and group significance
- Group validation techniques
- Multivariate analysis
- Congruence of techniques

Export

- Professional printing of reports and analyses
- Export as enhanced metafiles or bitmaps
- Customized reporting using scripts

Database sharing

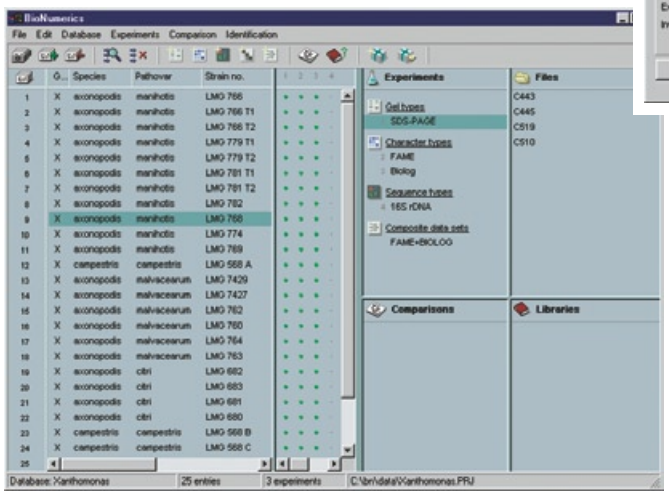
- Peer-to-peer data exchange
- Client-Server projects

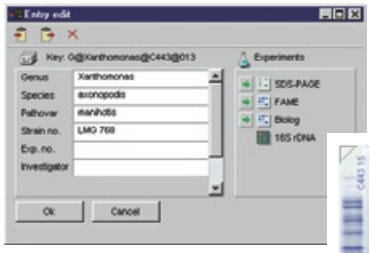
With the growing need to combine phenotypic, genotypic, electrophoresis and sequence information in the biosciences, researchers are confronted with two basic problems: (1) storage and retrieval of large amounts of information of very diverse nature, and (2) exploration, analysis and comparison of the organisms or samples based upon all available data. Specific software exists for the analysis of gels, phenodata, DNA arrays or sequences on the one hand, and powerful databases exist for storage of large amounts of data on the other hand. Likewise, general-purpose mathematics packages allow the analysis and comparison of numerical information. All of these applications are pieces of a complicated puzzle called bioinformatics. There is an emerging need for an integrative software package that combines input and processing of all possible biological experiments with a powerful central database and conglomerative exploration, grouping and comparison tools. BioNumerics perfectly addresses this need, expressed by research and diagnostic laboratories of any size involved in comparative analysis of biological data. The software has gained worldwide recognition by daily use in many research sites, including universities, hospitals and public health centers, food, drug and pharmaceutical industries, and a wide range of federal and private laboratories involved in typing, quality control, screening, testing, breeding, etc.

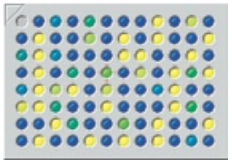
Concepts

The uniqueness of BioNumerics consists in the combination of a relational database with analysis modules for all existing biological data types. Besides bringing a powerful, multi-user database environment with unlimited possibilities in terms of size capacity and input sources, the same software offers the most advanced tools that exist for the analysis of gels and fingerprints, character arrays, and sequences. The biological entities of the database, i.e. the database entries, can be bacterial or viral strains, animals, plants, fungi, tissue, or any other organic samples for which fingerprint patterns can be obtained. The concept of the database allows various experiments of different nature to be defined for the same entry. As a result, multiple experimental data can be explored and compared among the entries studied, and groupings or identifications can be obtained for any combination of database entries and experiments available.

The experimental data can be subdivided into five classes, which include all possible experiment types employed to express relationships in biology: (1) 1-D electrophoresis patterns, called fingerprint types, (2) character-based data, called character types, (3) DNA, RNA, and protein sequences, called sequence types, (4) 2-D electrophoresis gels, called 2-D gel types, and (5) matrix types. Within each of these experimental classes, the user can create custom experiment types with particular settings. The five experiment classes are the basis of the modular subdivision of the BioNumerics software.



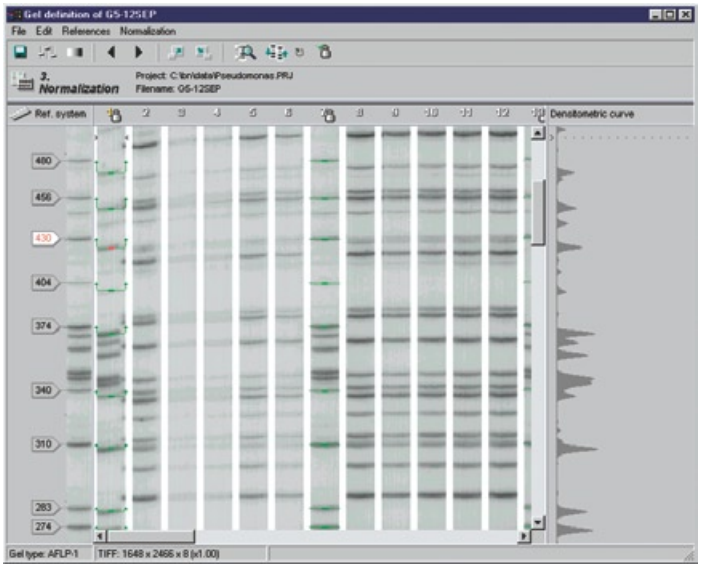




Fingerprint types

Any densitometric record seen as a profile of peaks or bands can be considered as a fingerprint type. Examples are electrophoresis patterns, gas chromatography or HPLC profiles, spectrophotometric curves, etc. For example, the user can create a Pulsed Field Gel Electrophoresis (PFGE) experiment type with specific settings such as reference marker, MW regression, staining color, band search settings, band matching tolerance, similarity coefficient, clustering method, etc. Electrophoresis is an important component in studying relationships in biology; therefore, comprehensive tools for preprocessing electrophoresis fingerprints are incorporated into BioNumerics. These tools include reading graphical and densitometric file formats from image files and automated sequencers, lane finding, normalization (alignment of patterns), band finding and quantification, band matching, etc.

The quality and completeness of electrophoresis fingerprint analysis in BioNumerics is illustrated by the fact that the famous GelCompar II software, with all its functions and possibilities, is entirely contained in BioNumerics' Fingerprint types application.

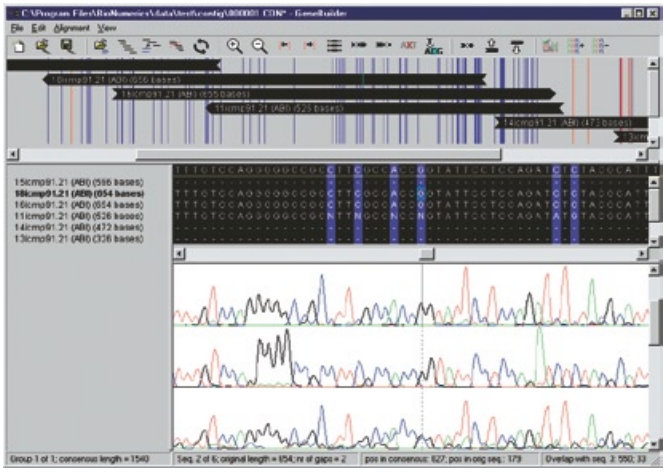


Character types

Using the character types, it is possible to define any array of named characters, binary or continuous, with fixed or undefined length. The size of a character type in BioNumerics can range from one single character (e.g. a morphological feature) to microarray experiments of many thousands of gene expression values. Further adjustable features include the range of the characters, the number of digits, the color scale, similarity coefficients used for comparison, etc. Examples of character types include fatty acid profiles, metabolic assimilation or enzyme activity test panels such as API, Biolog, Vitek, antibiotics resistance profiles, morphological and biochemical features, microarrays and gene chips, etc. BioNumerics' powerful script language and ODBC functions allow direct import of data from external databases or from text-oriented files or Excel spreadsheets.

Sequence types

Within the sequence types, the user can enter sequences of nucleic acids (DNA and RNA) and amino acids. BioNumerics recognizes widely used sequence file formats such as EMBL, GenBank, and Fasta, with the possibility to import user-selected header tags as information fields. In addition, BioNumerics' powerful sequence assembler module,



GeneBuilder, allows direct import of raw chromatogram files from automated sequencers. GeneBuilder has both an excellent alignment engine and a smart, user-friendly interface. Complete gene assembly projects with aligned chromatograms can be saved into projects and popped up with a single mouse-click on a sequence stored in the BioNumerics database. Each sequence type in BioNumerics can be stored with its own reference sequence, and with specific alignment and clustering settings.

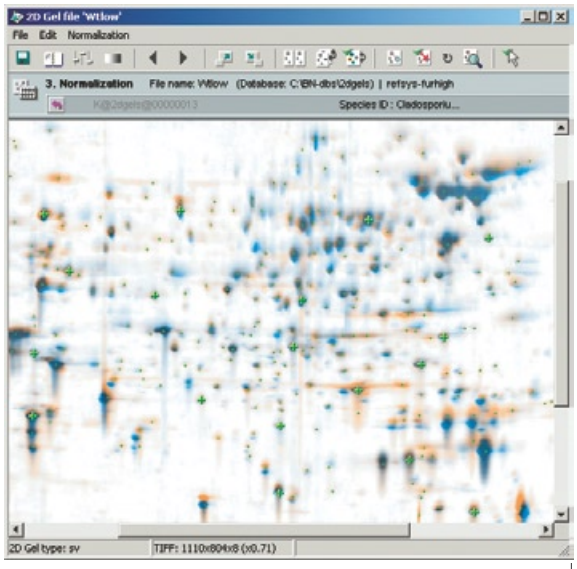
BioNumerics offers probably the finest sequence alignment and clustering tools that currently exist for PCs. It combines clustering of thousands of nucleotide or protein sequences of almost unlimited length with multiple alignment and display of homology matrices. The user is able to specify all alignment and homology parameters in the successive steps needed for a multiple alignment, such as open gap cost, unit gap cost, treatment of unknown bases, and the clustering algorithm. Multiple alignments associated with dendrograms can be edited manually, specific subsequences can be searched for, and complete comparisons can be saved to disk.

In addition to well-established alignment algorithms described in the literature, the software contains extremely fast and reliable algorithms elaborated at Applied Maths.

The software offers a wide range of phylogenetic clustering techniques and various tools for the estimation of the significance and reliability of clusters, which are discussed below under the Analysis and comparison functions.

2-D gel types

The 2-D gel analysis module in BioNumerics, also available as a standalone software package 2Dexpert, is a high level application for complete analysis and databasing of 2-D gels. Applied Maths' experience in image analysis has been fully exploited here to achieve more reliable automatic gel alignments than ever obtained so far.



In addition, interactive overlay images with gels shown in different colors allow the user to manually correct normalizations and detect unique and common spots at a glance. BioNumerics allows protein spots from 2-D gels to be identified and stored in the database. As such, 2-D gel information can be analyzed in an unparalleled way, using all the available querying, clustering, identification, and ordination techniques available in BioNumerics.

Matrix types

With matrix types, it is possible to import external similarity or distance matrices, providing similarity between entries revealed directly by the technique, or by other software. These matrices can be linked to the database entries in BioNumerics and they are used in conjunction with other information to obtain classifications and identifications. A typical example of a native matrix type is a table of DNA homology values.

Composite data sets

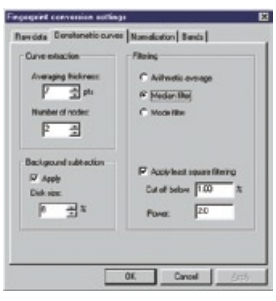
Each single database entry, a strain, organism, or sample, can have several experiments of different types linked to it. Linking an experiment to an entry is as easy as dragging the pointer to the corresponding database entry.

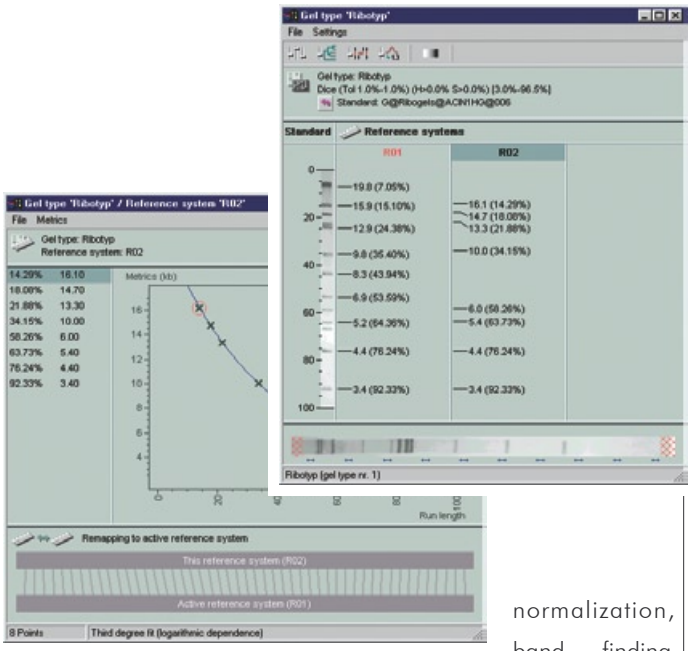
For example, a bacterial strain in the database could be characterized by an RFLP pattern, a fatty acid profile, a 16S rDNA sequence, an antibiotics resistance profile, etc. A plant cultivar or variety could have an AFLP pattern and a microarray linked to it. There is virtually no limit to the number and variety of experiments that can be linked to a single object under study.

When more than one experiment is available for a set of entries, it may be interesting to generate an overall table of characters, which includes all the characters of the available experiments, or a selection made by the user. The result is a sixth class of experiments, the so-called composite data set. A composite data set may include character types, sequences, 1-D or 2-D gels and matrices, and can be used for cluster analysis or identification, just like a single experiment type.

Analysis of 1-D electrophoresis fingerprints

Within the Fingerprint types module, BioNumerics offers an impressive gel analysis system, which benefits from Applied Maths' more than nine years experience and leadership in the field of electrophoresis typing. It can be safely stated that BioNumerics is the most advanced fingerprint analysis software that currently exists, in terms of reliability, user-friendliness, possibilities, and capacity. The software handles 8-bit, 12-bit, and 16-bit TIFF files as well as densitometric curves from sequencers, scanners, and spectrophotometers. Convenient wizards enable the user to define new fingerprint types and choose optimal settings for normalization, resolution, background subtraction, smoothing, band finding, etc. The whole process of analyzing a gel, starting with track definition,

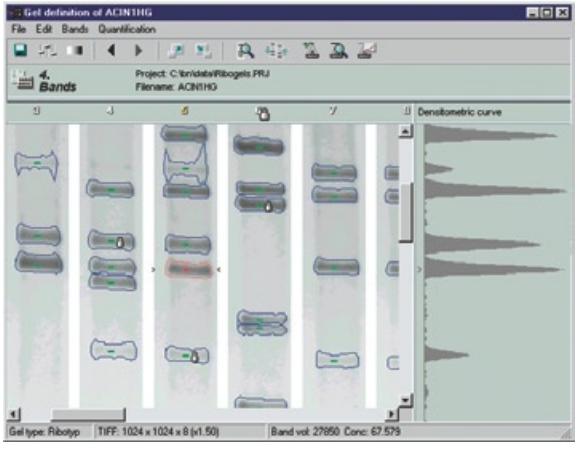




normalization, band finding, and ending with quantification, is contained in a powerful wizard, allowing the user to re-edit the gel at any stage without losing any editing done previously.

The user can always recall any step performed on a gel, an important feature in environments with GLP requirements. In addition, the software records a log file, rigorously keeping track of any change made.

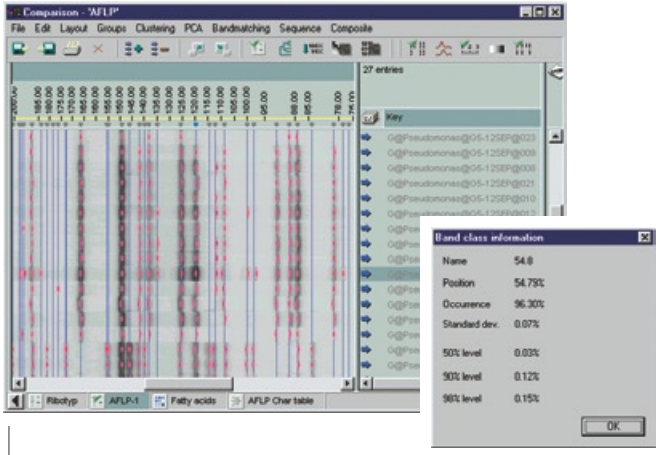
Reference bands used for alignment can be given a name or a molecular weight, which is used by the software to calculate the MW regression. The full information of reference bands used for the normalization of a specific type of gel, is called a reference system. An important innovation in gel analysis is the real-time normalization of pattern images and band positions, based on the reference system. This concept not only permits re-editing as described above, but also makes it possible to automatically remap one reference system into any other, as soon as there is sufficient overlap in the reference bands used, or the MW range of both systems. The result is that pattern images from gels run under different conditions and using different reference markers, can be displayed next to each other, with equimolecular bands perfectly aligned, as if the patterns were coming from the same gel.



Reliable quantification of bands or peaks is often a requirement in molecular research, in genetic breeding, and for quantitative comparisons. Here again, BioNumerics offers an uncompromising solution, for both 1-D densitometric records and 2-D scannings. One-dimensional peaks are modeled by best-fitting Gaussian curves, and 2-D images can be quantified by determining the contours of the bands. Calibration based upon regression of known band volumes results in a reliable estimate of concentration.

Unlimited zoom functions in all images, and convenient buttons, tool tips, and floating menus make the processing of gels highly surveyable and give the user easy and quick access to the wealth of advanced features available in BioNumerics.

Numerous other features such as spot removal, 2-D and 1-D background subtraction, filtering, spectral analysis, alignment distortion bars, definition of uncertain bands, optimization & tolerance statistics have made BioNumerics and GelCompar II the absolute standard for fingerprint analysis in environments where speed, volume and reliability are critical issues.

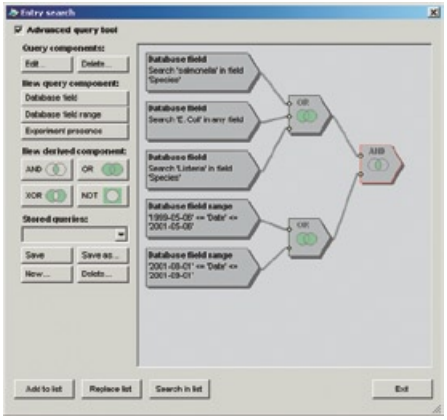


As a last important feature, the script language in BioNumerics allows any action involved in gel processing to be executed from a script, which makes it possible to introduce various levels of automation in the gel analysis procedure. In environments where large numbers of standardized gels are run, this feature forms an invaluable basis for low cost – high throughput routine analysis.

Databasing

BioNumerics has its own database engine, which is specifically designed for storage of diverse types of biodata. The capacity of the database is very high, while

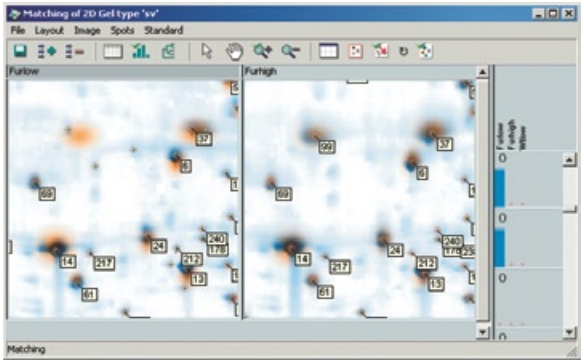
the maintenance is easy, which makes it the perfect choice for local single-user analysis. For lab-wide usage of centrally maintained databases, the Connected Database option is recommended, in which BioNumerics manages relational databases using an external database engine, such as Oracle, SQL Server or Access. The advantages of Connected Databases are true multi-user access, several levels of protection offered by the database engine, unlimited size capacity, transferability, advanced backup tools, and the possibility of implanting BioNumerics in an existing database environment or LIMS.



One of the highly appreciated database features that characterize BioNumerics, is the advanced querying tool.

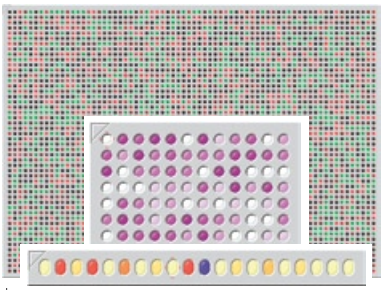
Query components can be created based upon database fields, ranges of fields, availability of experiments, presence of bands or characters, character values, subsequences, etc. These components can be combined using logical operators such as AND, NOT, OR, XOR, giving rise to complex queries that are nicely represented in a smart interactive diagram. Queries can be saved to be reused and modified at any time.

Every biological experiment, including gel patterns, densitometric curves, carbohydrate assimilation panels, antibiotic resistance profiles, blots, microarrays, 2-D gels and sequences, can instantly be visualized with a single mouse-click and comparisons between experiments can be shown.



Analysis and comparison

In addition to the five experiment type modules, BioNumerics offers three comparison type modules: (i) Cluster Analysis and phylogeny, (ii) Non-hierarchic



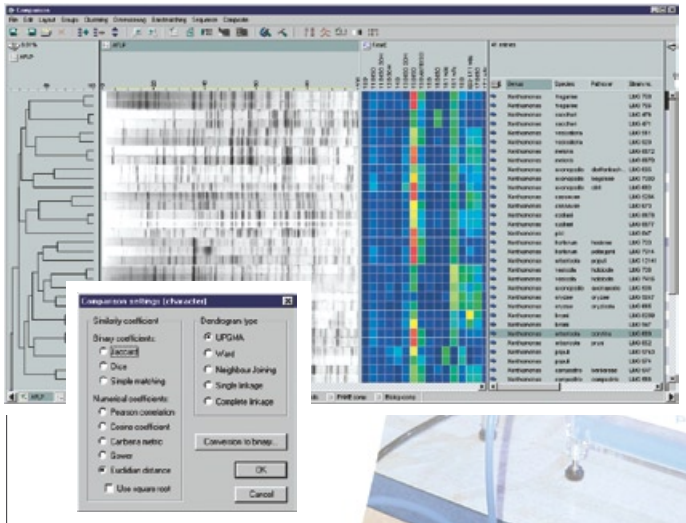
grouping techniques, also called dimensioning techniques, and (iii) Libraries and identification. Each of these modules is very comprehensive in terms of mathematical and technical possibilities, so that only their most important features can be highlighted in the following paragraphs.

Cluster Analysis and Phylogeny

Since the availability of computers to biologists, cluster analysis has been a fundamental tool in classification, screening, typing, and epidemiology. Putting together the concepts of a relational database, multiple experiment linkage, and a range of powerful clustering algorithms has resulted in a clustering module with unique capabilities in BioNumerics.

The Comparison window. This crucial window in BioNumerics presents a comprehensive overview of all available experiments for a selection of entries and enables the user to show and compare any combination of experiments.

Similarity or distance matrices and dendrograms can be calculated for any selected experiment, and the obtained groupings can be compared with patterns or characters obtained from other experiments. A variety of similarity and distance coefficients and clustering methods are available, in order to provide the most appropriate clustering for all data types.



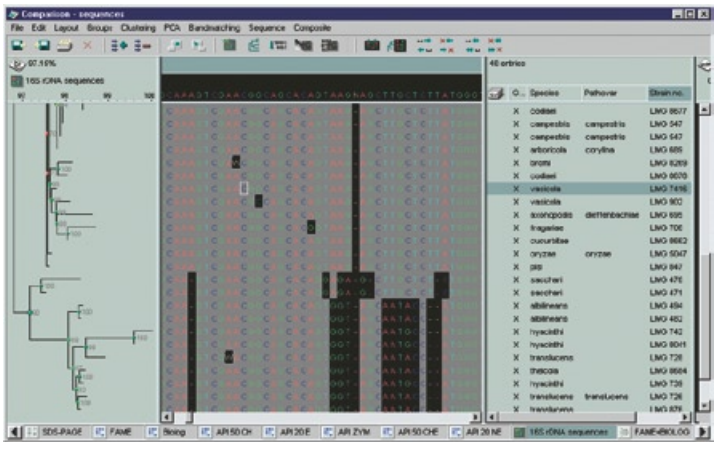
Composite cluster analysis.

Composite clusterings can be generated from selected combinations of experiments, and various methods can be used to obtain a combined dendrogram. Similarities can be adopted from the individual experiments and averaged by user-defined weights, or weights determined by the program, based upon the amount of information available in each experiment. Alternatively, all characters from the individual experiments can be pooled to form one global data set, which can be clustered. Advanced mathematical algorithms allow the calculation of a consensus similarity matrix and dendrogram based upon individual matrices from different experiments.

Dendrogram functions. The number of database entries that can be clustered into a single dendrogram is virtually unlimited. Several advanced editing tools such as swapping and abridging of branches, zooming in or out, rerooting of trees, changing shading intervals of similarity matrices, make the interpretation of large cluster analyses easier.

Incremental clustering. The unique "incremental clustering" algorithm allows batches of entries to be pasted, or deleted from existing dendrograms without having to recalculate the entire similarity matrix. BioNumerics automatically updates the existing matrix and rebuilds the dendrogram accordingly, so that adding or deleting entries becomes a matter of seconds instead of minutes or hours.

Special attention has been paid to the incremental construction of multiple alignments of sequences. Upon deletion or addition of sequences, alignments are rebuilt by reconstructing partial consensus sequences from the branches that have changed down to the root. In order to maintain existing alignments edited by the user, groups of sequences can be locked. This implies that, if new sequences are added, these locked groups will be aligned to them as one consensus.



Dendrogram significance tools. Several statistical methods are available for evaluating the confidence level of a global tree, and of each individual branch. These methods include the standard deviation and cophenetic correlation at each branching level and the root, bootstrap analysis at each branching level of a (pseudo) rooted tree, and the Jackknife method. Partitioning methods provide an alternative way to discovering group structures in complex data sets.

Clustering of characters. Not only can entries be clustered based upon their common and different characters, but also characters can be simultaneously clustered based upon the entries they share. This approach results in a transversal clustering, a combined view in which both database entries and characters are clustered, and which allows the user to easily reveal the characters that determine and distinguish groups of related entries.

Phylogenetic inference. In addition to pair-wise clustering techniques such as UPGMA, Ward, Single Linkage, Complete Linkage and Neighbor Joining, BioNumerics offers true phylogenetic clustering algorithms based upon evolutionary optimization criteria. These include the Generalized Maximum Parsimony method and the Maximum Likelihood algorithm. Parsimony can be combined with bootstrap analysis whereas maximum likelihood offers the Likelihood Ratio Test. Both methods result in an unrooted "seaweed" dendrogram, which can be converted into a "pseudo-rooted" tree after assignment of a root. To correct phylogenetic distance scaling, the Jukes & Cantor or Kimura 2 parameter correction factors can be chosen.

› Dimensioning techniques

Under dimensioning techniques, we classify all techniques that place the entries in a two- or more dimensional space, rather than imposing a hierarchical, bifurcating structure like a dendrogram.

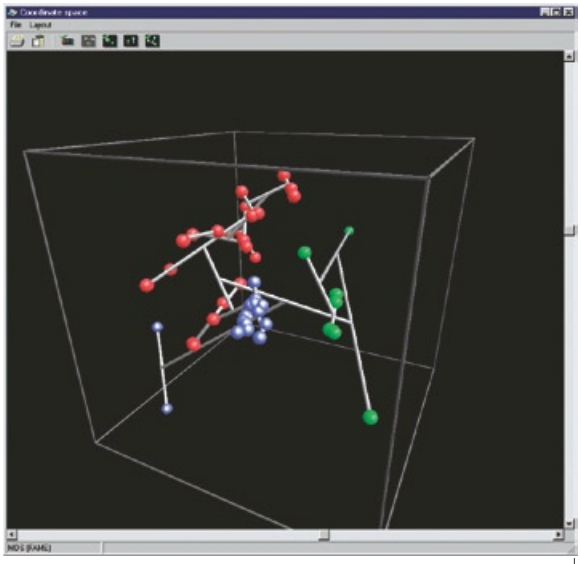
Principal Components Analysis (PCA). This technique starts directly from a character table to obtain groupings in a multi-dimensional space. Any combination of axes can be displayed in two- or three dimensions.

Multi-Dimensional Scaling (MDS). Rather than starting from the data set, MDS uses the similarity matrix as input, which has the advantage over PCA that it can be applied directly to banding patterns. The MDS algorithm iteratively optimizes the distances between the entries in the MDS space according to the similarity values of the matrix.

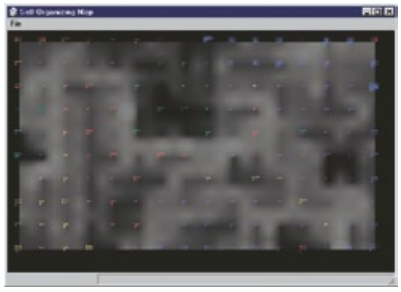
The advanced presentation modes of both PCA and MDS produce fascinating three-dimensional graphs in an X-Y-Z coordinate system, which can rotate in real time to enhance

the perception of the spatial structures. An innovative feature in BioNumerics is the function for connecting the spatially scattered entries according to the branching of a corres-ponding dendrogram. This not only provides an excellent comparison between the two grouping approaches, but also improves the understanding and the perception of the obtained groups. New entries can be placed in an existing PCA or MDS for identification.

Discriminant Analysis and MANOVA. These advanced statistical analysis methods allow the relation between groups of entries and characters to be discovered, and the significance of such groups to be determined. The groups can be clusters derived from a dendrogram, or any user-defined selections of entries (e.g., by origin, serotype, …).

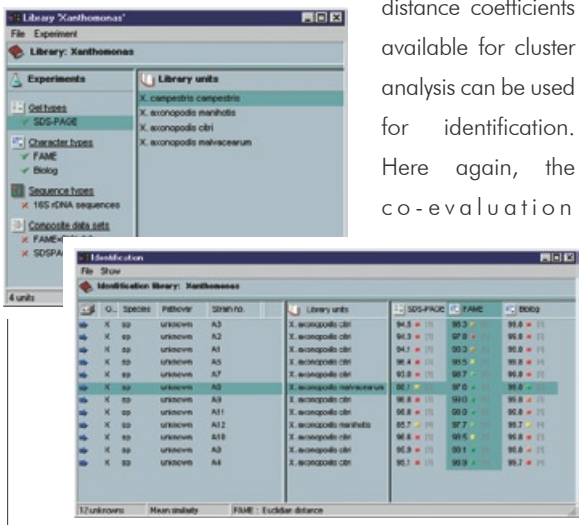


Self-Organizing Maps (SOM). Basically being a type of neural network, a SOM is able to place many thousands of entries in a two-dimensional representation, a map, according to overall relatedness. For complex data sets with large numbers of entries, SOM analysis is to be preferred over traditional clustering. An interesting option of a SOM is that unknown entries can be placed in an existing map, which offers a quick and reliable identification tool. BioNumerics was the first software to apply this exciting technique to biological relatedness study and for identification.

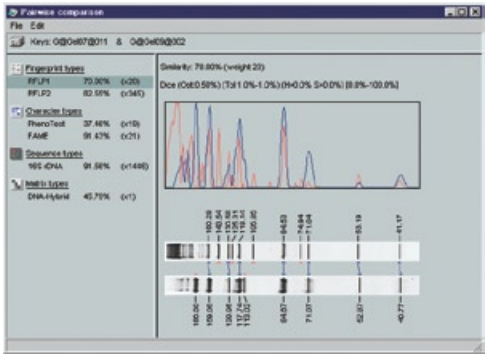


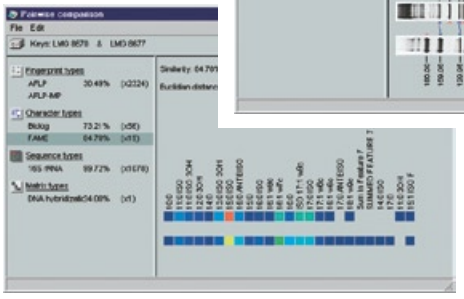
› Libraries and identification

The possibility of identifying unknown organisms based upon various available experiments is also a big step forward realized in BioNumerics. The basis for identification is a library. An identification library is a collection of units, each of which consists of one or more entries of the same taxon or group. The identification of unknown organisms depends on the similarity to the available library units. The same range of similarity and distance coefficients available for cluster analysis can be used for identification. Here again, the co-evaluation



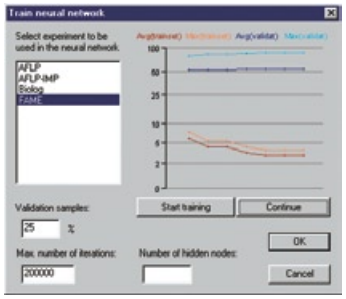
of different available data sets for library entries and/or unknown entries leads to more faithful consensus identifications. In addition to the individual experiment types, the user may create any composite data set based upon different experiment types, and use this set for identification. The software automatically searches for all corresponding data between a library entry and the unknown entry, to use them in the calculations of similarity whenever available.





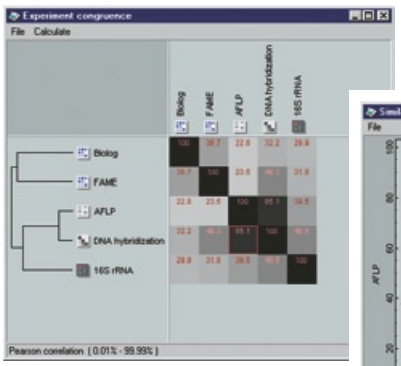
An identification report lists the global identification as well as the identifications obtained by all individual data sets. Mathematical and statistical methods allow the estimation of the reliability and the relevance of each identification case.

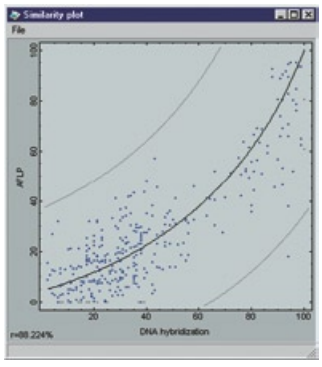
A detailed pairwise comparison can be obtained between any two entries from the database, which lists all the experiments that both entries share, together with the percentage similarity. With a simple mouse-click on the experiment type, the gelstrips, character sets, or aligned sequences of both entries are shown together. As an interesting alternative to classical similarity-based identification, BioNumerics allows a neural network to be generated for each experiment type. For large databases containing groups that are difficult to distinguish, a neural network can be the quickest and most reliable identification tool.



Analysis of congruence between techniques

When comparisons are made between groupings based upon different techniques, the question arises to what extent there exists any congruence between these different techniques. Another interesting aim is to find which technique is the closest to the consensus classification, since this technique will in general be the most reliable for identifying the organisms or samples under study. This is another analytical tool offered by BioNumerics: similarity matrices obtained from different techniques are compared in a pairwise manner by comparing corresponding similarity values by either Kendall's Tau coefficient or the product-moment correlation. This results in a congruence matrix, expressing the global similarity or congruence between different techniques. This matrix in turn can be clustered into a dendrogram, now grouping techniques according to congruence.





Pairwise comparisons between any two techniques are obtained by plotting the corresponding similarity values in an X-Y diagram. Such plots are very useful both to reveal the taxonomic level or depth of one technique compared to another: it shows whether one technique is discriminative at a lower or higher level than another technique and provides insight into the limitations and benefits of each technique in building identification strategies.

Database sharing

Today, the exchange of information among different laboratories is of the utmost importance in the scientific world. The need to exchange fingerprint information is not an exception to this rule, and has become particularly urgent in epidemiological research and surveillance networks. BioNumerics offers a powerful solution to this important issue with its integrated Database Sharing Tools. The key component behind the exchange of data is the so-called Bundle: a packed file that contains a selection of Bio-Numerics database entries (e.g. bacterial strains), along with any selection of experiments and information fields specified by the user.

Peer-to-peer data exchange. The Database Sharing Tools allow BioNumerics users to exchange information at a peer-to-peer level by simply making a selection of database entries, and clicking the information fields and experiments to be contained in the bundle. Received bundles can be directly analyzed together with other database entries. BioNumerics automatically recognizes which experiments are compatible.

Client-Server approach. BioNumerics' advanced client-server system is the perfect solution for collaborative research projects, networks, and private initiatives where central databases are made available to a restricted or unrestricted number of client users. Each BioNumerics software package that contains the Database Sharing Tools comes as a client version, which can connect and communicate with a BioNumerics Server using TCP/IP. A direct connection is established between the Server and the Client allowing uploading and downloading of database entries, interactive querying, and automatic identification of profiles uploaded by the client. Using the script language both at client and server site, the most sophisticated implementations can be designed. Examples include automatic creation and broadcasting of reports and notices, or automatic alerts of members in surveillance networks.

Modular structure

BioNumerics consists of 9 different modules in total, of which 5 modules are related to the different experimental applications that can be analyzed (application modules), and 4 modules constitute the different analysis tools that the software offers (analysis modules).

- 5 application modules: Fingerprint types, Character types, Sequence types, 2-D gel types, and Matrix types.
- 4 analysis modules: Cluster analysis, Identification & Libraries, Dimensioning techniques, and Database Sharing Tools.

The full BioNumerics functionality is physically contained in the same program unit, which guarantees perfect integration of the modules and easy co-evaluation of different data sets and analyses. For example, a selection of entries highlighted on a dendrogram (Cluster analysis module) becomes also highlighted on a PCA (Dimensioning techniques module) and in the database.

Any or all of the application and analysis modules can be combined with each other. At least one application module is required to operate the software.

Compatibility

Import of fingerprints: Accepts uncompressed gel images as 8-bit, 12-bit, and 16-bit TIFF files generated by any flatbed scanner, video camera or fluoro-imaging system. Processing and normalization of multichannel chromatogram files from automated sequencers (ABI, Beckman). Direct import of absorbance profiles from a variety of scanners, sequencers and RiboPrinter system. Import of processed gel data as MW or curves. Scriptable import of densitometric records available in any known format.

Import of character data: RGB channel quantification of grid-based character data scanned as 8-bit to 24-bit TIFF images, such as microplates, phenotypic test panels, DNA arrays, etc. Import of character data from Excel spreadsheets or databases using ODBC. Scripts available for import of most common phenotypic test panels and fatty acid profiles. Scriptable import of any character array available in any format.

Import of sequences: Processing and contig construction of multichannel chromatogram files from automated sequencers (ABI, Beckman) and text or binary files. Compatible with EMBL, GenBank, and Fasta sequence formats. Import of DNA, RNA, and protein sequences. Import of aligned sequences. Scriptable import of less common file formats.

Import of database information: Import of information fields from any text file type using scripts. Direct import from external databases using ODBC.

Printing and export

Highly professional print reports in color or grayscale. Each graphical or text-oriented print job can be copied to the clipboard for import in other Windows software, or can be saved as bitmap file with adjustable resolution. Creation of custom graphics or text report possible using scripts.

Script language

Powerful script language to realize tasks like importing data from files or databases, fingerprint processing, task automation, exporting BioNumerics data to other databases, creating customized graphics and text reports, manipulation of database fields, manipulation of experimental data, performing complex queries, creating specific analysis tools, etc.

Computer requirements

PC-compatible with Pentium CPU or better, 64 MB RAM or more, 65K color graphics or better and Windows 98, Windows NT 4.0 or higher. CD-ROM and mouse are required. Recommended screen resolution: 1024x768 or higher.

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