GelCompar® II
TODAY'S FOREMOST SOFTWARE FOR THE ANALYSIS OF BANDING PATTERNS AND FINGERPRINTS

www.applied-maths.com
Ever since its first release in 1991, GelCompar has been a household name in comparative analysis of electrophoresis patterns. Along with the success and expansion of electrophoresis and other fingerprinting techniques in various application fields, GelCompar has been continuously updated and extended to be compatible with the latest developments and the ever growing amounts of information. As a part of its long history, the software was fully redesigned to share all the advanced concepts and components with the renowned BioNumerics software, resulting in a more powerful yet user-friendly software package, GelCompar II. The database concept in GelCompar II makes it possible to link multiple fingerprints to single organisms or samples studied, and to generate consensus groupings and identifications. In addition, GelCompar II offers an impressive list of unique advanced features, including phylogenetic and ordination algorithms, population modelling, group verification methods, database quality control techniques, history recording, SQL compatibility, script language, database sharing tools, client-server database exchange over Internet, just to name a few. With thousands of units sold in the most prestigious research institutes worldwide and being used in numerous international epidemiological typing and surveillance networks, GelCompar II and its parent BioNumerics proudly bear the gold standard for the analysis of fingerprint patterns.

Some of the main features of GelCompar II:

- Easy processing of all types of gels and densitometric fingerprint files, including bitmap files of any size, multichannel sequencer chromatograms, spectrophotometric, chromatographic and HPLC records, MALDI and SELDI profiles, etc.
- Comprehensive image and curve enhancement tools such as spot removal, 2-D background subtraction, noise and background filtering, tone and spectral analysis.
- Advanced automatic lane finding, normalization, and band/peak searching with settings to fit any type of fingerprint.
- Reliable quantitation of sizes and concentrations.
- Organism-oriented multi-level relational database with almost unlimited capacity and advanced query tools.
- Ability to link multiple fingerprints to organisms/samples.
- Generation of band scoring tables and analysis of polymorphic bands.
- Comprehensive cluster analysis, parsimony and population genetics on single or combined fingerprint types.
- Principal Components Analysis and Multidimensional Scaling.
- Creation of 2-D and 3-D bar graphs, contingency tables, and scatterplots from database fields and band characteristics; various significance tests.
- Cluster and group verification statistics based on bootstrap, standard deviation, and Jackknife.
- Fast and accurate identification based on similarity coefficients or neural networks.
- Peer-to-peer data exchange and client-server connection with BioNumerics Server.
- Script language for automation of tasks, import and export, generation of custom reports, customization of interface, etc.
- Rich base of free plug-ins and script-based tools for specialized analysis needs.

GelCompar II is a modular package, consisting of the Basic Software and four modules for specific applications.

- **GelCompar Basic Software**: Processing of images and densitometric profiles including normalization and band searching. Full database functionality and display of normalized profiles. Calculation of molecular sizes and quantitation of bands or peaks. Quantitation and comparison of bands between groups of patterns and generation of binary or quantitative scoring tables. Analysis of unique, common, polymorphic, and discriminative bands.


- **Identification and Library Manager**: Open and unlimited multi-library system for identification. Various identification methods with statistical interpretation, including Neural Networks.

- **Dimensioning Techniques**: Non-hierarchic grouping by PCA and MDS with 2D and 3D presentation modes. Creation of plots, graphs, and contingency tables.

- **Database Sharing Tools**: Possibility to create databases in Oracle, Microsoft Access, and SQL Server. Exchange of database and pattern information using XML. Connectivity with BioNumerics Server.

All GelCompar II functionality is contained in one single application, which guarantees perfect integration of the modules, a consistent user-interface, and no tedious switching between programs.
CONCEPTS

GelCompar II relies on two basic concepts: the database entries and the experimental data. The relational database consists of entries which correspond to the individual organisms or samples under study: bacterial or viral strains, animals, plants, fungi, tissue, or any other organic samples for which fingerprint patterns can be obtained. Each database entry is characterized by a unique key, assigned either automatically by the software or manually by the user, and by additional user-defined information fields.

The experimental data are the results of the electrophoresis or other fingerprint experiments that are used to estimate the resemblance between the organisms or samples. The user can create custom fingerprint types, each with particular settings for normalization, band detection, quantitation, clustering, etc.

The concept of the database allows various fingerprint experiments of different nature to be defined for the same entry. As a result, multiple fingerprint 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 “Fingerprint” type of data

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, MALDI and SELDI profiles, etc.

Within a single GelCompar II database, the user can create many different fingerprint experiment types, each with specific settings such as type of input files, reference marker set, background subtraction strength, smoothing and filtering settings, peak search and shoulder sensitivity, MW regression, stain, peak/band matching tolerance, similarity coefficient, clustering method, etc. Since electrophoresis is an important component in studying relationships in biology, comprehensive tools for preprocessing electrophoresis fingerprints are incorporated in GelCompar II. These tools include the readout of various graphical and densitometric file formats, automatic lane-finding, normalization (alignment of patterns), peak finding, quantitation, peak/band matching, etc.

Image processing and normalization

The successful generation of databases of electrophoresis patterns depends on the quality and reliability of preprocessing of the images and/or profiles. Preprocessing involves four important steps: (i) image enhancement (for bitmap images), including smoothing, contrast and brightness optimization, spot removal, background subtraction, etc.; (ii) densitogram enhancement, involving background subtraction, averaging, noise filtering, spike removal, smoothing, etc.; (iii) normalization, which means the correction of profiles to match each other over different runs, and (iv) band detection. GelCompar II offers very sophisticated pattern preprocessing and normalization tools drawn upon Applied Maths’ experience and leadership of over 16 years in the field of electrophoresis typing. The software handles 8-bit and 16-bit image files as well as densitometric formats, e.g. from automated sequencers, scanners, or spectrophotometers. Convenient wizards enable the user to define new fingerprint types and choose optimal settings for normalization, resolution, background subtraction, smoothing, band finding etc. Advanced functions such as tone curve analysis, spectral analysis and distortion analysis provide the more experienced user with invaluable information about gel quality and optimal settings. Gels or lanes can be displayed as high-contrast embossed images or as fascinating three-dimensional images, which can rotate in real-time. The whole preprocessing of a gel (or set of fingerprints), start-
ing with track definition, normalization, band finding, and ending with quantitation, is contained in a powerful and user-friendly stepwise procedure, 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 keeps track of any changes made to files, which are accurately recorded to a log file.

Normalization of patterns is based on external reference patterns or internal reference bands (multi-dye chromatogram files). Reference bands used for alignment can be assigned a name or a molecular weight, which is used by the software to calculate the MW regression. The full set of reference positions and the associated MW regression curve is saved as a reference system. Powerful pattern recognition systems allow automated and accurate assignment of bands to reference positions, and a complex non-linear interpolation algorithm ensures the most reliable normalization possible for all types of gels and non-linear distortion. In case of very distorted gels, the user can intervene at any time and apply manual corrections. To make the editing process easier, the software offers a friendly multi-level undo function.

An important innovation in gel analysis is the real-time normalization of pattern images and band/peak positions. This concept not only permits re-editing as described before, but also makes it possible to compare gels from different origin and using different reference markers, as soon as there is sufficient overlap in the size range of both systems.

Defining bands or peaks on patterns can often be a critical and time-consuming task. GelCompar II offers accurate band/peak search algorithms that are amenable to all types of patterns through a number of adjustable parameters. The software allows bands/peaks within certain intensity thresholds to be marked as uncertain, in which case they are neither considered as a match, nor as a mismatch in comparisons. For techniques where the band/peak intensity differs in function of the size (e.g. EtBr stained gels such as PFGE), a peak intensity regression can be created based upon processed database patterns. The software uses the obtained regression to define peaks with much higher accuracy.

Zoom-sliders in all images, convenient buttons, tool tips, floating menus, and multilevel undo/redo features make the processing of gels easy and highly surveyable and give the user easy and quick access to the wealth of advanced features available. Numerous advanced features, such as spectral analysis, definition of uncertain bands, optimization and tolerance statistics, have made GelCompar II the absolute standard for fingerprint analysis in environments where speed and volume as well as reliability are critical issues.

As a new important feature, the script language in GelCompar II 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 generated, this feature forms an invaluable basis for low cost - high throughput routine analysis.
Button bars can be docked or removed as desired. Individual buttons can be added or removed.

Database panels can be docked conveniently or removed if not used. Panels can also be tabbed with other panels to optimize display usage.

Clear icons in menus and buttons aid quick access to frequently used tools.

All database components are described by useful information fields that can be shown or hidden, queried, sorted, etc. The user can define custom fields.

Enter information and edit fields directly in the main window.
The backbone of GelCompar II is a powerful relational database, specifically designed for storing and retrieving biological data. By default, the software will create Microsoft™ Access Databases, which are suitable for most purposes, including occasional multi-user access. For high-volume databasing, lab-wide access, permission control, automatic backup etc., GelCompar II will also manage a number of more professional database engines such as Oracle™ and SQL Server™. The rich and flexible database structure allows information to be added at numerous levels. For example, an organism can have its own descriptive information fields (up to 150), and can also have a number of attachments associated with it. These include text files, images, Word, Excel and PDF files, and HTML/XML files or URLs. Fingerprint files and fingerprints can have their own information fields.

A fingerprint pattern, including its gelstrip, densitometric curve, and band assignments, can instantly be visualized from virtually any application, and detailed comparisons between patterns can be shown. Experiments linked to the organism, for example the gel pattern and the gel file in which the pattern occurs, can have their own descriptive information fields. Even comparisons, subsets, libraries, and other objects can have associated information fields.

One of the highly appreciated database features that characterize GelCompar II, is the advanced querying tool. Query components can be created based upon database fields, ranges of fields, dates, availability of a pattern belonging to a specific fingerprint type, and presence of bands with specific size and/or quantity. 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.

GelCompar II offers four major classes of comparison techniques: (i) polymorphism analysis based upon global band matching tables, (ii) cluster analysis and phylogenetic methods, (iii) non-hierarchic grouping methods (also called dimensioning techniques) and statistics, and (iv) identification and the construction of libraries.

In addition, the design of the relational database makes it possible to analyze and compare multiple patterns per entry in a so-called composite data set. This approach allows for the combined analysis of patterns obtained with different restriction enzymes, primers or probes, which results in higher resolution clusterings and identifications.

Quantitation and polymorphism analysis

Reliable quantitation of bands or peaks is often a requirement in molecular research, in genetic breeding, and for quantitative comparisons.

Here again, GelCompar II offers reliable solutions, both for densitometric records and two-dimensional scannings. One-dimensional peaks are approached by best-fitting Gaussian curves, and two-dimensional images can be quantified by determining the contours of the bands. Calibration is based upon non-linear regression of known band volumes, resulting in the most accurate estimation of quantity.

The possibility to generate global band scoring tables from selections of multiple patterns is particularly useful for the identification of molecular markers, for example for co-dominant scoring analysis in breeding experiments. This application has been thoroughly elaborated in GelCompar II, with the possibility to automatically create band classes according to user-specified toler-
fingerprints. Similarity matrices and dendrograms can be calculated from 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 clusterings for all data types.

In addition to conventional algorithms such as UPGMA, Single Linkage, Complete Linkage, Ward, and Neighbor Joining, the program offers phylogenetic inference tools based on generalized parsimony.

- **Composite clustering**. 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 number of band instances available in each trace-to-trace comparison. Alternatively, all characters from the individual band matching tables can be pooled to form one global data set, which can be clustered.

- **Dendrogram functions**. There is almost no limit to the number of database entries that can be clustered into a single dendrogram. Several advanced editing tools make the interpretation of large cluster analyses easier: swapping of branches, rerooting of trees, changing shading intervals of similarity matrices.

---

**Cluster Analysis and Phylogeny**

Since the availability of computers to biologists, cluster analysis has been a very important tool in classification, screening, typing, and epidemiology. Putting together the concepts of a relational database, linkage of multiple experiments, and a range of powerful clustering algorithms has resulted in the unique clustering module available in GelCompar II.

- **The Comparison window**. The so-called Comparison window in GelCompar II presents a comprehensive overview of all available experiments for a selection of entries and enables the user to show and compare any combination of images of
**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. GelCompar II automatically updates the existing matrix and rebuilds the dendrogram accordingly, so that adding or deleting entries to/from a clustering of thousands of entries takes only a few seconds.

**Dendrogram significance tools.** Several statistical methods are available for evaluating the confidence level of a global tree, and for each individual branch. These methods include the standard deviation and the cophenetic correlation at each branching level including the root, and bootstrap analysis. The Jackknife method can be used to test any set of groups defined by the user.

**Clustering of characters.** Not only can entries be clustered based upon their common and different characters in a band matching table via a so-called Q-matrix, but characters can be simultaneously clustered based upon the entries they share (R-matrix). This approach results in 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 organisms.

**Phylogenetic inference.** In addition to pair-wise clustering techniques such as UPGMA, Ward, Single Linkage, Complete Linkage and Neighbor Joining, GelCompar II offers a true phylogenetic clustering algorithm based upon the Generalized Maximum Parsimony method. Parsimony can be combined with bootstrap analysis and results in an unrooted “seaweed” dendrogram. The unrooted tree can be converted into a “pseudo-rooted” tree after assignment of a root.

**Minimum Spanning Trees.** Whereas parsimony and maximum likelihood techniques are suitable for inferring deeper phylogenetic relationships, the Minimum Spanning Tree (MST) algorithm allows short-term divergence and micro-evolution in populations to be reconstructed based upon sampled data. The MST technique as implemented in GelCompar II is an excellent tool for analyzing binary or categorical data such as derived from VNTR or AFLP fingerprints. The MST interface offers great interaction with the database and other techniques and is the ideal platform for e.g. plotting epidemic divergence against other factors such as geographical distribution, date of sampling, serotypes, etc.
Dimensioning techniques and statistics

Two general multivariate analysis techniques are available in GelCompar II: Principal Components Analysis (PCA) and Multi-Dimensional Scaling (MDS). Whereas PCA starts from a character table to obtain groupings in a multi-dimensional space, MDS uses a similarity matrix obtained after trace-to-trace comparisons using one of the available similarity or distance coefficients. The advantage of the latter technique is that it can be directly applied to banding patterns, whereas PCA can only be applied to global band matching tables.

Although less commonly used than grouping methods that infer dendrograms, PCA and MDS are at least as attractive and faithful in discriminating a small number of groups or closely related organisms. Both techniques can be applied to data sets from single experiments or on any combination of experiments.

The advanced PCA and MDS presentation modes available in GelCompar II produce fascinating three-dimensional graphs in an X-Y-Z coordinate system, which can rotate in real-time to enhance perception of the spatial structures. A useful feature in GelCompar II is the function for connecting the spatially scattered entries according to the branching of a corresponding 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. The Dimensioning tools module of GelCompar II also offers a nice set of statistical and graph tools, allowing the user to create 2D and 3D bar graphs, contingency tables, and scatterplots from database fields, band characteristics, and combinations of both.

Libraries and identification

Besides clustering, identification is another analysis tool in biology. The possibility to identify unknown organisms or samples based upon a combination of different fingerprint types boosts the reliability of critical identification applications.

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 experiments, the user may create any composite data set based upon different fingerprint types, and include this set for identification. The software automatically searches for all corresponding data between any library entry and the unknown entry, to use them in the calculations of consensus similarity when available. An identification report lists the global identification, as well as the identifications obtained using all individual data sets. Statistical methods allow the estimation of the reliability and the statistical
As an interesting alternative to classical similarity-based identification, GelCompar II allows neural networks to be generated for each fingerprint type. For large databases containing groups that are difficult to distinguish, a neural network can be the quickest and at the same time the most reliable identification tool. A detailed pairwise comparison can be obtained between any two entries from the database. All the experiments that both entries share are listed together with the percentages similarity. With a simple mouse-click, the gels strips and fully detailed comparisons are shown for the selected experiment, together with the densitometric curves, molecular weights and matching or non-matching bands.

**Database sharing and client-server tools**

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.

GelCompar II offers a total solution to this important issue with its integrated Database Sharing Tools. The key component behind the exchange is the integrated library of XML functions. Using the XML export tools, a packed XML file can be created, containing a set of GelCompar II database entries (e.g. bacterial strains), along with any selection of experiment results and information fields, as defined by the user. Using the XML import functionality, the XML files can be extracted to reappear as native, editable database entries, along with all the biological information that was saved with the selection.

- **Peer-to-peer data exchange.** The database exchange tools in GelCompar II allow 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 XML file. Received files can be opened and analyzed directly together with other database entries. GelCompar II automatically recognizes what fingerprint types are compatible and applies a molecular weight remapping if necessary.

- **Client-Server approach.** Applied Maths’ advanced client-server system, built around the BioNumerics Server, 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 GelCompar II 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.

- **Geographical mapping.** In many research projects, especially epidemiological, biological data is closely linked to geographical data. The Database Sharing Tools enable a Geo plugin to be installed, offering a simple yet powerful way to map the results from queries, comparisons, identifications etc. on geographical maps. Geographical information with database entries can be provided as city names, postal or zip codes, or geographical coordinates. Entries can be plotted individually or as stacked bar graphs or pie charts, using different colors according to groups defined in the database. The powerful geographic tools of
Google™ Maps and versatile search, select, and query interface of GelCompar II together make the Geo plugin a very useful and interactive asset.

Plugin tools

Although GelCompar II is a versatile and comprehensive platform for the analysis and databasing of any type of fingerprint data, a number of applications are too specific to be provided in a generic environment. This is the case for specific import and export tools, but also for a number of cutting-edge techniques that require continuous updating of the analysis tools to keep up with the latest developments. Therefore, most technique-oriented functionality has been enabled as Plugin applications. These plugins are well-documented in separate manuals and are officially supported by Applied Maths. GelCompar II currently offers free plugins for import and export, XML-based exchange, geographical mapping, and a wide variety of extra functionality for the database, fingerprint processing, dendrograms tools, and reporting.

Compatibility

- **Import of fingerprints:** Accepts gel images as 8-bit and 16-bit bitmap files in TIFF, JPG, BMP, GIF, or PNG format, generated by any imaging system. Direct import and processing of multichannel chromatogram files from automated sequencers (Applied Biosystems, Beckman, Amersham). Import of absorbance and densitometry profiles from a variety of scanners, sequencers and automated system (electrophoresis, spectrophotometry, HPLC, mass spectrometry, MALDI, SELDI, etc. Import of processed densitometric data as peak MW, RF, height and/or surface tables. Scriptable import of any densitometric or peak table record available in text format.

- **Import of database information:** Import of information fields from any text file type, spreadsheet or database using the import plugin. Direct link with SQL and ODBC compatible databases. Easy import of uncommon formats using the script language.

- **Printing and export:** Professional print reports in color or grayscale. Each graphical print job can be copied to the clipboard as high-resolution metafile for import in other Windows software, or can be saved as bitmap file with adjustable resolution. Text-oriented reports and listings can be exported as tab-delimited files or copied to the clipboard. Creation of custom graphics or text reports is possible using scripts.

- **Script language:** Powerful script language to realize tasks like importing data from files or databases, automated import and processing of fingerprints, exporting data, creating customized graphics and text reports, manipulation of database fields, manipulation of experimental data, performing complex queries, creating specific analysis tools, building forms and interfaces, etc.