Bio-Structural Classification Database

Summer 2007 Program on the Geometry and Statistics of Shape Spaces.
Statistical and Applied Mathematical Sciences Institute (SAMSI), Research Triangle Park, NC 27709-4006. July 9, 2007.

Juan B. Gutierrez, Christian Laing, Monica K. Hurdal
Florida State University, Department of Mathematics, Tallahassee, FL 32306-4510

1 Introduction

Information about biological structures (e.g. brains, proteins, nucleic acids, etc) is usually comprised of data such as numbers, characters, images, and structures of soft complexity, i.e. data organized according to some logic that is not or explicit or self-evident. While there is a vast variety of types of bio-structures, most of them can be represented for data mining purposes with a reduced set of mathematical objects that require large amounts of data for proper representation. Efficiently managing these large datasets becomes a major computational challenge, which we address with the Bio-Structural Classification Database (**BSCD**). It is an information system that facilitates mathematical research by encapsulating all the complexity related to data management.

2 Bio-Structures

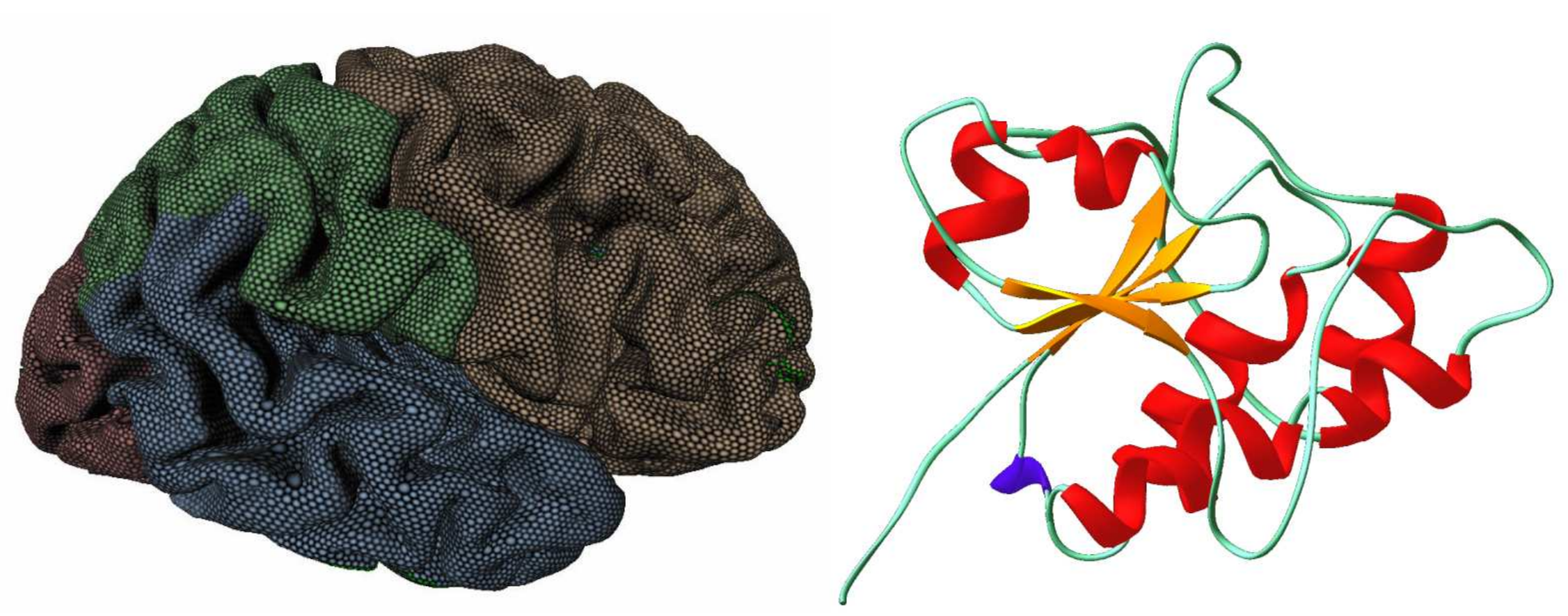


Figure 1: *Different bio-structures (brain and protein) that can be represented with the same type of mathematical object: graphs.*

The bio-structures that are processed by the **BSCD** are graphs which can be used to represent surfaces and/or polygonal curves, such as those shown in figure 1. The representation of a surface is given by a triangulated mesh, S, described in terms of the set of vertices V in the surface S and the set of edges E in the mesh. Thus, $S = (V, E)$. The paths on the surface are curves in space that form a family Ω of polygonal curves in $\mathbb{R}^3$. Each curve, $\gamma \in \Omega$, has intrinsic measures of its geometry, useful to build an n-dimensional feature vector $\mathbf{F}$.

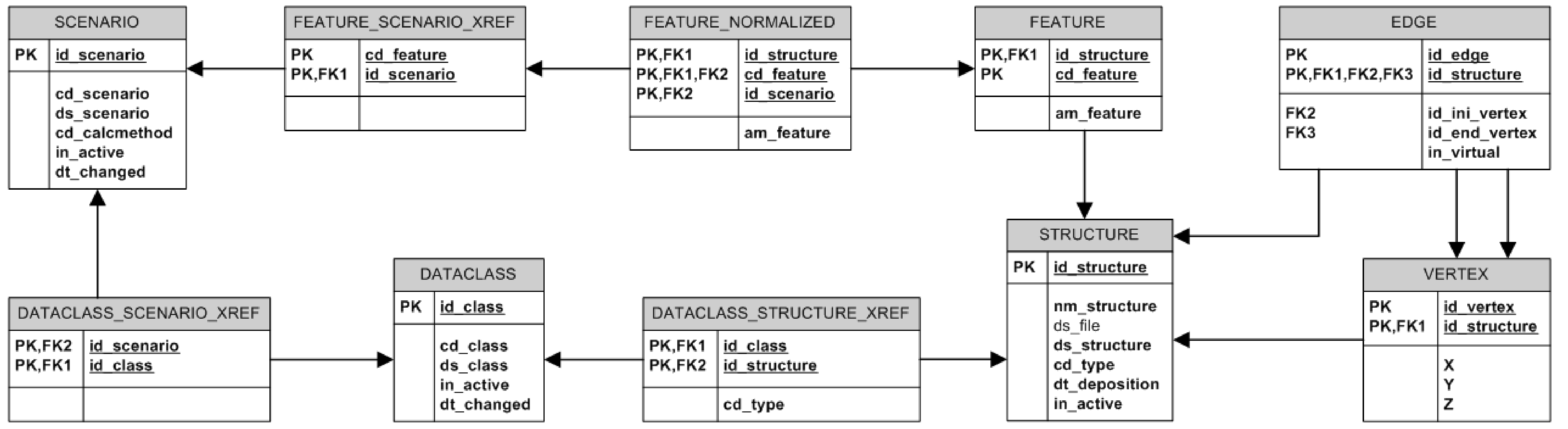
3 Pattern Classification

Given a partition of Ω into a finite set of classes $C = \{w_1, w_2, \ldots, w_c\}$, it is possible to construct a discriminant function $g_i(\mathbf{F})$, $i = 1, 2, \ldots, c$ that assigns $\mathbf{F}$ to a class w_i if $g_i(\mathbf{F}) < g_j(\mathbf{F})$ for all $j \neq i$. The classification problem requires the feature vector $\mathbf{F}$ to be comparable across bio-structures.

The n-dimensional feature vector can be projected into $\mathbb{R}^2$ or $\mathbb{R}^3$ and analyzed with Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA). PCA and MDA differ mathematically in what they maximize. MDA maximizes the difference between the classes, whereas PCA maximizes the variance in all the classes. It is usually the case in a PCA and/or MDA representation that when a simple visual inspection can identify clusters, the classification algorithms yield favorable results [1, 3, 2]. Features calculated in the system are analyzed in Weka, an open-source collection of machine learning algorithms for data mining [3].

4 Results

The information system was built using .NET technology, SQL Server 2005, and Windows 2003 Server. The database design is shown in figure 2. The **BSCD** could handle up to 32,767 database files of 32 TB each for a total of 1,048,516 TB. The system can run in grid mode, each grid with maximum 32 GB of RAM, and 32 processors, with no practical limit for the grid size. Screenshots of the web interface are shown in figure 3. The **BSCD** is available at http://www.bioclassification.org.



SCENARIO
PK | id_scenario
cd_scenario, ds_scenario, cd_calcmethod, in_active, dt_changed

FEATURE_SCENARIO_XREF
PK | cd_feature
PK,FK1 | id_scenario

FEATURE_NORMALIZED
PK,FK1 | id_structure
PK,FK1,FK2 | cd_feature
PK,FK2 | id_scenario
am_feature

FEATURE
PK,FK1 | id_structure
PK | cd_feature
am_feature

EDGE
PK | id_edge
PK,FK1,FK2,FK3 | id_structure
FK2 | id_ini_vertex
FK3 | id_end_vertex
in_virtual

STRUCTURE
PK | id_structure
nm_structure, ds_file, ds_structure, cd_type, dt_deposition, in_active

VERTEX
PK | id_vertex
PK,FK1 | id_structure
X, Y, Z

DATACLASS_SCENARIO_XREF
PK,FK2 | id_scenario
PK,FK1 | id_class

DATACLASS
PK | id_class
cd_class, ds_class, in_active, dt_changed

DATACLASS_STRUCTURE_XREF
PK,FK1 | id_class
PK,FK2 | id_structure
cd_type

Figure 2: *Entity-relationship diagram. The relational database was normalized to the third normal form.*

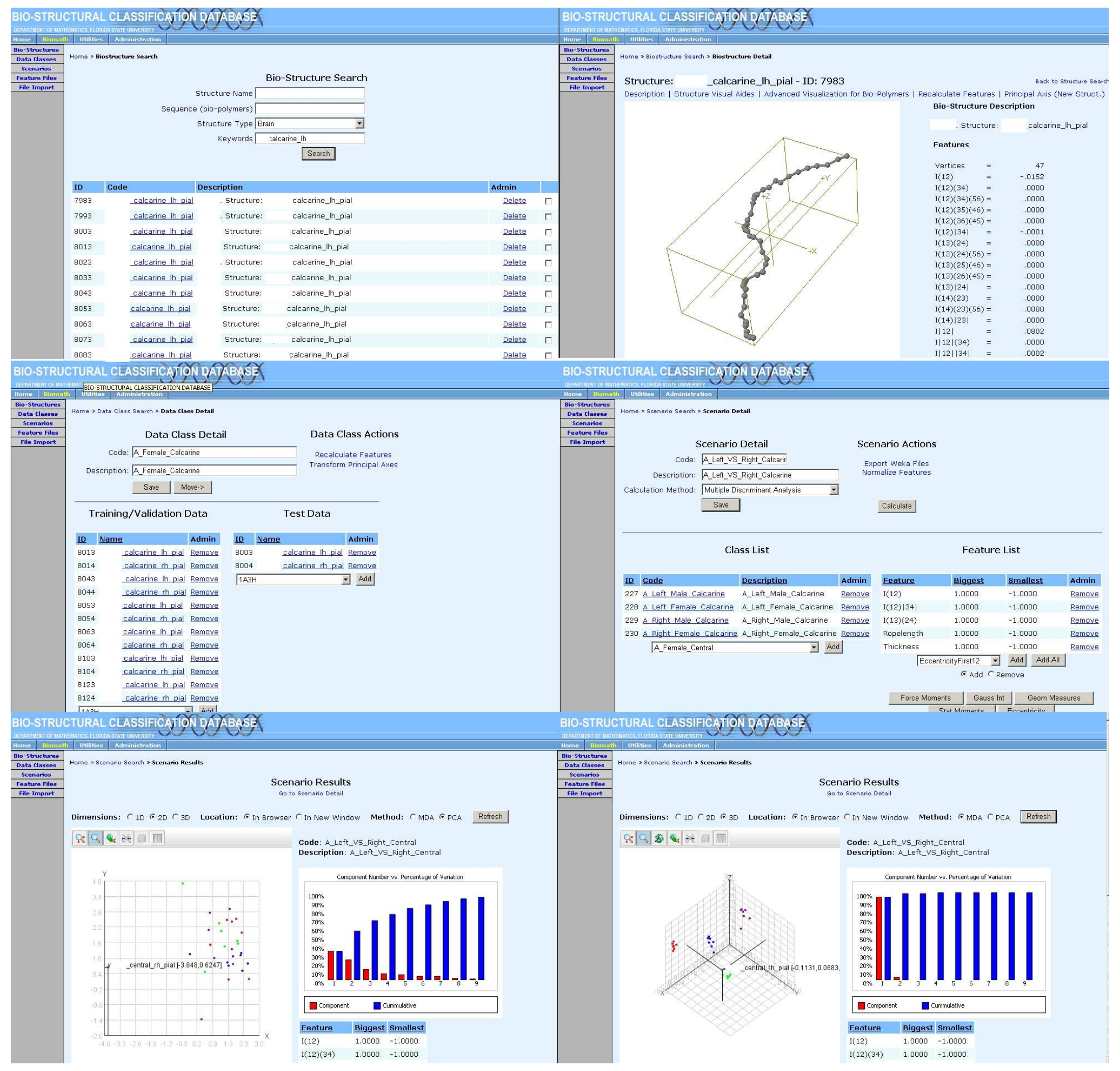


Figure 3: *Screenshots of the web interface.*

Bio-Structures are grouped in classes. Any structure can belong to many classes. Classes are grouped by classification scenarios. Any class can belong to many scenarios. Classification method and features are selected uniquely per scenario. The user-friendly interface allows users to remove and add classes to the scenarios and structures to the classes on the fly, expediting the exploratory phase of feature research.

5 Conclusions

We built a web front-end to a high performance relational database for pattern classification of bio-structures that is capable of managing massive data. This system produces different types of feature vectors, mainly higher order moments and knot invariants based on Gauss integrals, and then applies a battery of pattern classification algorithms. This high-performance information system has been successfully used to classify biological data.

Acknowledgments

This work is supported in part by the Program in Mathematics and Molecular Biology (PMMB) through a Burroughs Welcome Fund Interfaces Grant and *Centro Avanzado de Investigación en Inteligencia Artificial* (CAVIIAR - Advanced Research Center in Artificial Intelligence). We thank Debbie A. Smith for her use of the **BSCD** and the valuable feedback she provided.

References

[1] R. O. Duda, D. G. Stork, and P. E. Hart. *Pattern Classification*. John Wiley, Inc., Indianapolis, IN, 1999.

[2] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 2002.

[3] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.