A Granular Hierarchical Approach to Biophysical and Biochemical Evaluations of Transgenic Modifications

Moshe A. Gadish, David K. Chiu, Cecil W. Forsberg  
Department of Computing and Information Science  
University of Guelph  
Guelph, Ontario, Canada  
mgadish@uoguelph.ca

Serguei P. Golovan  
Department of Animal and Food Science  
University of Delaware  
Newark, DE, United States

Abstract—A novel information-theoretical systems approach for the comparative analysis of multivariable organismal level data is proposed. The system structure is hierarchical, represented as multiple levels of granulation. It is used to model biological transgenic modifications measured on an organism. Experimental evaluation showed the approach being successful in detecting known biochemical and biophysical variable interactions.

Keywords—bioinformatics; systems biology; granular computing; mutual information; transgenic organism;

I. INTRODUCTION

With the advances in biotechnology involving the creation of transgenic organisms, investigation of the relationships among biological expression components has been a major goal of bioinformatics and systems biology. The process of linking genotypic and phenotypic attributes will play a vital role in understanding biological processes, cellular and organism responses under novel conditions, such as disease or genetic modifications (i.e., transgene). While number of such analysis was performed, they often concentrate on identification of significantly affected gene and proteins yet the analysis of significantly affected interactions is often missing. Also the integration and analysis of different types of high resolution data from microarray, proteomic and metabolomic experiments is often difficult exercise.

The objective of this study is to develop new system of information-theoretic analysis for multivariable data analysis. Biophysical and biochemical data collected from transgenic pigs that utilize plant phosphorus more efficiently [1,14] and ordinary (non-transgenic) pigs were used to create the model. The system is described as a hierarchy of multiple levels of granulation: a base layer representing the raw data; a transformation layer that transformed continuous data into a common discrete type; a network layer of pair-wise relationships, based on statistically evaluation of significant expected mutual information; a convergent pattern layer indicating strong dependency that identifies high summation nodes of relationships. The patterns detected in these layers can then be used to perform additional analysis in comparing detected interactions between transgenic species and that of ordinary ones.

A. Systems Biology

Systems biology studies the interactions between the components of biological systems, and how these interactions give indications to the function and behavior of that system [2]. There are many approaches in systems biology for inferring biological networks of interactions, possibly using parametric or non-parametric modeling. Most research involving gene expression may focus on inferring interactions of gene expression, produced from microarray experiments. These gene interaction algorithms can be classified into several categories, including: clustering techniques; Bayesian networks [3]; dynamic modeling [4]; correlation [5], and mutual information [6] based approaches. The approach proposed here, is information-theoretic analysis. It employs mutual information to detect both linear and non-linear interactions between biological variables from synthetic levels of possibly genotypic and phenotypic variables, regardless of their data type. Mutual information is used as a more generalized approach for correlation. It has several advantages. Firstly, it can detect all interactions including both linear and non-linear relationships. Secondly, it requires no parametric modelling, as is the case in Bayesian network, dynamic control theory and multiple regression approaches. Because our approach integrates variables of different types combining them into a hierarchical and network structure of different scale, it is most suitable to present a systems view for analysis of multiple gene expression values; or chemical concentrations. The system proposed here, will be presented as a comparison to granular computing.

B. Granular Computing

Granular computing explores the composition of parts, their interrelationships, and connections to the whole. A problem can be represented as a web of interacting and interrelated parts. In order to have understanding and derive a practical solution, it may be necessary to extract approximate structures that are tractable and easy to analyze. Granular computing exploits structures in terms of granules, levels, and hierarchies based on multilevel and multi-view
representations [11,12,13]. In our approach, the data are analogous to granules, units of information representing a part of a whole. A granule may be interpreted as one of the numerous small particles forming a larger unit. A granule may have two distinctive roles: a more refined representation of another broader granule, or a network family of granules considered as a whole. Granular structures provide a structural description of a system supported by the data analyzed. A system’s granular structure can then be described by the granules internal structure; the collective structure of a family of granules; and the hierarchical structure of a web of granules. The collective structure can be viewed as a level in hierarchy structure as well as an interconnected network of granules. The paper is structured as follows. In section II, we shall describe the approach, introducing the mathematical and statistical notations. Section III describes the experiments conducted in this study, comparing analysis of variable taken from a transgenic animal, as comparing to its ordinary one. Section IV provides a discussion of the experimental results. Finally, Section V presents the conclusions of this work.

II. METHODS OF ANALYSIS

Here we provide an overview of our approach. Data attributes are organized into various categories, forming a set. Each attribute can be considered as a variable where the corresponding value of that property at an instance is the outcome. In our representation scheme, the data attributes are initially transformed using a discretization schema for data that are continuous. Each variable is then compared to another, using only those pairs of observations that are evaluated to be statistically significant. The resulting value is the expected mutual information that is significant between each pair of variables, which in turn is normalized to yield a value of interdependency between 0 and 1. Our method then selects those variables that show strongest interdependency with others within the set, thus identifying those variables that form a convergent pattern. The final stage of the analysis compares detected interactions between classes or between that of transgenic species with ordinary ones (Fig 1).

A. Level 0: Base Information

For variables that have continuous value (such as concentration of minerals), a scheme is developed to discretize the variable. After discretization, each variable value is substituted with its corresponding calculated label for that property. Each variable is divided into n equal intervals,

\[
Interval = \frac{\text{max} - \text{min}}{n}
\]

where max and min are the maximum and minimum observed values for a variable, and Interval is the interval size. Each property then falls into one of the predefined n intervals. Observations that share similar values fall into the same interval and are assigned identical discrete values. This process is repeated for the entire data set, producing transformed observations for each variable, with a specified accuracy of discretization. This is a type of range discretization, where the data range of each interval is equal, and the number of data values in each interval varies according to the interval range.

B. Level 1: Systematic Reduction of Information

Biological data attributes can be organized into various genotypic and phenotypic categories, forming a set. Each phenotypic and genotypic property can be represented as a variable where the corresponding value of that property is the outcome. The set can be represented as \(X = (X_1, X_2, ..., X_m)\) where \(m\) is the number of variables, indicating the size of the data set. An instance of \(X\) is a realization denoted as \(x=(x_1, x_2, ..., x_m)\). Each \(x_j (1 \leq j \leq m)\) can take up an attribute value denoted as \(x_j = a_{jq}\). An attribute value \(a_{jq}\) is a value taken from an attribute value set \(E = \{a_{i}\} i=1, 2, ..., L\) where \(L\) is the number of possible values for the variable \(X_j\), or the cardinality of the set.

C. Level 2: Detection of Significant Pair-Wise Interactions

Expected mutual information is a measure of the statistical interdependence between two variables. The stronger the interdependence between the two variables, the larger is the expected mutual information between them. If the two variables are strictly independent, then the expected mutual information between them is zero [7,8]. It is defined as,

\[
I(X_i, X_j) = \sum_{j=1}^{L_j} \sum_{k=1}^{L_k} P(X_i^j, X_j^k) \log \frac{P(X_i^j, X_j^k)}{P(X_i^j)P(X_j^k)}
\]

It is important when calculating statistical interdependence to take into consideration their statistical significance, so that their correspondence is not due to chance, otherwise considerable error can be accumulated when summation is used to indicate a convergence of a pattern. The adjusted residual, is chosen for this purpose,
\[ d(e_{ik}^h) = \frac{z(e_{ik}^h)}{\sqrt{v(e_{ik}^h)}} \] (3)

Expected mutual information as defined in Equation (2) subjected to the selections from the statistical test, using the statistics defined in Equation (3),

\[ I^*(X, X_i) = I(X, X_i) | d(e_{ik}^h) > N_{\alpha} \] (4)

where \( N_{\alpha} \) is the table threshold value with a statistical significance level \( \alpha \). Significant expected mutual information, can be normalized to produce values between 0 and 1 by dividing it to Shannon entropy involving only those outcomes selected. Shannon entropy can be denoted as,

\[ H^*(X, X_i) = \sum_{j} \sum_{h} P(X_j^i, X_h^i) \log P(X_j^i, X_h^i) | d(e_{ik}^h) > N_{\alpha} \] (5)

The normalized expected mutual information (denoted as SEMI) based on the selected significant events, can now be defined as,

\[ R^*(X, X_i) = \frac{I^*(X, X_i)}{H^*(X, X_i)} \] (6)

D. Level 3: Detecting Strong Interactions

To evaluate the total amount of interdependency expressed on a given variable induced by the detection of \( R^* \), it can be calculated as,

\[ MR(X_i) = \sum_{k=1}^{l_k} R^*(X_i, X_i) \] (7)

E. Hierarchical Comparative Analysis

Hierarchical structures can be observed in many natural, artificial and abstract systems. A hierarchy links the components into a whole, providing a multi-level and multi-resolution representation of the system.

Hierarchical representation reflects the orderliness, control, and stability of a such a system. Systems that share similar base levels can then be compared on multiple levels of granulation. Computation with hierarchies involves the comparison from within the same level across different classes. Consider a pair of systems representing two classes of animals (for example, transgenic vs. ordinary animals). A comparison between two 2nd level granule structures can be done by comparing the significant expected mutual information between the corresponding interaction granules. For example,

\[ R^\text{Difference}(X_i, X_i) = R^\text{Class}_1(X_i, X_i) - R^\text{Class}_2(X_i, X_i) \] (8)

If we are comparing a transgenic animal (class 1) with an ordinary one (class 2), this computation at this level will find all these interactions between the two classes. A comparison between two 3rd level granule structures, will determine the differences with respect to the convergent patterns, such as

\[ MR^\text{Difference}(X_i) = MR^\text{Class}_1(X_i) - MR^\text{Class}_2(X_i) \] (9)

III. EXPERIMENTAL DESIGN

The Enviropig™ animal, developed at the University of Guelph, was chosen to evaluate the proposed method. The Enviropig™ is a transgenic pig expressing phytase in the salivary glands that is able to utilize plant phosphorus more efficiently than ordinary (non-transgenic) pigs leading to significant decrease in phosphorus pollution [1,14]. The ability to identify negative side effects of transgene expression is generally considered to be extremely important for improving genetically modified food safety. The Enviropig™ has been compared with ordinary pigs, with respect to gene expression, chemical composition and meat quality attributes. There is a need for a system representation for the Enviropig™, setting up the framework for discovering significant differences between the transgenic and ordinary pig with respect to both genotype and phenotype attributes.

Data were collected from both the transgenic pig line and the ordinary one. They consist of several groups of measurements. Among them are data, gathered from male and female pigs, as well as several biochemical and biophysical components in the meat tissues. The data is organized as follows: chemical composition attributes; hematology and blood biochemistry attributes; meat cuts quality attributes and proteomics.

The approach was initially tested on a reduced subset. Chemical composition data from the muscle tissue was used. The data consists of 55 chemical composition attributes as sampled from 48 pigs: 24 ordinary Yorkshire pigs (12 males and 12 females) and 24 (12 males and 12 females) transgenic Yorkshire pigs. The proposed method was then tested. The data was discretized into fixed 4 equal-width intervals, Due to the low sample size, the data was tested at a statistical significance level of 99% as well as the more common 95%.

Because the method is designed to handle a large amount of data involving various aspects of the transgenic animal, significant expected mutual information is visualized in three ways. A matrix plot, shows the value of significant expected mutual information (normalized), \( R^* \), between every pair of variables in a set. It is a two dimensional grayscale plot depicting interactions between any pair of variables. The strength of the relationship is inversely proportional to the intensity of the color (the darker the color, the stronger the interdependency between the variables). A cumulative line plot visualizes the convergent cumulative significant expected mutual information, (denoted as MR), between a variable and the rest of the variables in the set. An interaction plot, as an undirected cyclic graph shows strong interactions (above a threshold computed from a selected level of confidence) between pairs of variables.

IV. RESULTS AND DISCUSSION

A. Comparing Ordinary to Transgenic Pigs

Initially the approach was tested on the entire sample set. The detected interactions produced the expected results (Fig. 2). Interactions between Fat and Kilocalories variables are expected since they are both energy measures. Detected interactions between fatty acids are also explainable: Fatty acids C18:2, C18:3 are Omega-6 and cis polyunsaturated fatty acids; Fatty acid C18:1 is a monounsaturated fatty acid;
Fatty acid C22:4 is an Omega-3 polyunsaturated fatty acid [9]. Interactions between Magnesium, Potassium and Phosphorus are expected: in grains and beans about two-thirds of plant phosphorus is present in the form of insoluble phytate particles with high concentration of potassium, magnesium, and calcium [15]. Magnesium is also necessary for the function of the sodium/potassium pump [10].

Next, we divided the data set into two unique subsets: the first subset contains only transgenic samples (male and female); while the second subset contained only ordinary samples (male and female). The data was analyzed for interactions, and then visualized. Figures 3 and 4 visualize detected interactions for transgenic and ordinary samples respectively, while Table I describes the difference between them. Detected interactions are represented by four types of patterns: peaks, representing variables with few interactions; gaps, representing variables with very few and low interactions; single, strong pair-wise interactions; and cluster, several variables with strong mutual interactions (Fig. 5).

Clearly, there are more and stronger (i.e., darker cells) interactions for ordinary pigs. The data is organized into three categories: gross composition (fat, protein, moisture and ash), 43 fatty acids that make up the fat followed by 12 minerals that are components of the ash fraction. Common to them. Detected interactions are represented by four types of variables with strong mutual interactions (Fig. 5).

Comparing ordinary to transgenic pig interactions (Table I) revealed strong difference of interactions in C18:3 Linolenic and C20:4 Arachidonic fatty acids. Interactions were observed to have similar pattern, sharing many peaks and gaps at the same positions, while differing at the level of interaction strength, being consistently stronger in the ordinary pigs.

<table>
<thead>
<tr>
<th>Variables (position, value)</th>
<th>Pattern (variable name, position, value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cis-Polysaturated Fatty Acids (6, 5.358)</td>
<td>Peak (C18:3 Linolenic, 27, 0.647)</td>
</tr>
<tr>
<td>Omega-6 Polysaturated Fatty Acids (9, 4.7)</td>
<td>Peak (C18:3 Linolenic, 27, 0.563)</td>
</tr>
<tr>
<td>Trans-Fatty Acids (11, 6.178)</td>
<td>Peak</td>
</tr>
<tr>
<td>C8 Caprylic (14, 4.307)</td>
<td>Peak</td>
</tr>
<tr>
<td>C18:3 Linolenic (27, 7.249)</td>
<td>Peak (cis-Polysaturated Fatty Acids, 27, 0.647; Omega-6 Polysaturated Fatty Acids, 9, 0.563; C18:2 Linoleic, 24, 0.437; C20:4 Arachidonic, 33, 0.554)</td>
</tr>
<tr>
<td>C20:4 Arachidonic (33, 7.045)</td>
<td>Peak (C18:3 Linolenic, 27, 0.554)</td>
</tr>
<tr>
<td>C22:6 DHA (40, 6.138)</td>
<td>Peak (C18:2trans, 25, 0.676)</td>
</tr>
<tr>
<td>Calcium (43, 4.669)</td>
<td>Peak</td>
</tr>
<tr>
<td>Magnesium (47, 5.246)</td>
<td>Peak (Phosphorus, 50, 0.521; Zinc, 54, 0.406)</td>
</tr>
<tr>
<td>Phosphorus (50, 5.947)</td>
<td>Peak (Magnesium, 47, 0.521)</td>
</tr>
<tr>
<td>C4 Butyric (12)</td>
<td>Gap</td>
</tr>
<tr>
<td>C18:4 Moraotic (28)</td>
<td>Gap</td>
</tr>
<tr>
<td>C20:5 W3 EPA (34)</td>
<td>Gap</td>
</tr>
<tr>
<td>C22:2 Docosadienoic (37)</td>
<td>Gap</td>
</tr>
<tr>
<td>C24:0 (41)</td>
<td>Gap</td>
</tr>
<tr>
<td>C18:2trans (25, 4.647)</td>
<td>Peak (C22:6 DHA, 40, 0.676)</td>
</tr>
</tbody>
</table>

Comparing ordinary to transgenic pig interactions (Table I) revealed strong difference of interactions in C18:3 Linolenic and C20:4 Arachidonic fatty acids. Interactions were observed to have similar pattern, sharing many peaks and gaps at the same positions, while differing at the level of interaction strength, being consistently stronger in the ordinary pigs.

Figure 2. Interaction plot for all samples, R* > 0.3. The strength of the relationship is color coded: blue (0.3, 0.4), red (0.4, 0.5), black (0.5, 0.75). Values for cumulative SEMI are displayed inside the brackets.

Legend
- 0.75 - 1
- 0.5 - 0.75
- 0.4 - 0.5
- 0.3 - 0.4
- cis-polyunsaturated FA
- Omega-6 Polysaturated FA
- C20:2 Eicosadienoic
- C18:3 Gamma Linolenic
- C18:1 Oleic
- C18:2 Linoleic
- C18:3 Linolenic
- C18:4 Moroctic
- C20:4 Arachidonic
- C22:6 DHA
- C24:1 Nervonic
- C22:5 Docosapentaenoic
- Other fatty acids
- Other minerals

Figure 3. Interaction plot for the entire transgenic line sample showing all detected interactions, using SEMI approach, where the strength of interactions is color coded. Common nodes are colorless.
that is non-parametric and data driven. It can detect interactions between component variables regardless of their linear or non-linear nature, scale, and data type (continuous or discrete). Preliminary experiments on chemical compositions taken from transgenic animals produced clear patterns of interactions, and differences of interactions between the pigs lines (transgenic vs. ordinary).

**ACKNOWLEDGMENT**

We would like to thank Natural Science and Engineering Research Council of Canada, Discovery Grant, and the Advanced Food and Materials Network (AFMNet) for funding our research.

**REFERENCES**


