Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Cluster Methodology Defines Archetype Sentinel Consomic Rats

Nancy Laning Sobczak[1.], George F. Corliss[1.], Martin A. Seitz[1.], Peter J. Tonellato[1.,2.]
Marquette University[1.]., Medical College of Wisconsin[2.], Milwaukee, Wisconsin

I. ABSTRACT

A clustering method was developed to identify rats demonstrating characteristics of hypertension ('sentinel' animals). Although all rats in the group do not demonstrate every symptom of hypertension, these 'archetype sentinels' do demonstrate essential characteristics of the disease. This method applies two computational techniques to create a mechanism for classifying individuals whose biomedical profile include a spectrum of disease related phenotypes. First, Fuzzy Cluster Means (FCM) is used to distinguish a very small group of archetype subjects from the general population. Archetypes are defined as those subjects that are most often classified correctly within a limited portion of the entire collection of biomedical phenotypes. The resulting archetype is then examined to determine which phenotypes best characterize the set. In this study, the defining set of phenotypes demonstrate essential symptoms of hypertension. Consequently, we consider the archetype set 'sentinel' animals for hypertension. The archetype sentinels are then used to train a Neural Network (NN) to determine the physiological characteristics of phenotyped subjects which are not the Archetype Sentinels.

Two inbred strains of rats are used: Brown Norway (BN) and the Brown Norway/Salt Sensitive Chromosome 20 (SS20BN). The physiological database contains a total of 63 phenotypes (41 renal, 22 cardiac). A total of 79 phenotyped rats were analyzed with the FCM method yielding 6 BN archetypes and 5 SS20BN archetype subjects characterized by a total of 39 phenotypes (18 renal and 21 cardiac.) These 11 archetype sentinels then were used as a neural network training set to classify the non-archetype sentinel subjects as either normal (BN) or hypertensive (SS20BN). Of all rats tested, 10 of 11 BN and 10 of 10 SS20BN rats were properly classified. Overall, 95% of the rats were classified correctly, with one false positive result.

Results demonstrate that the FCM method can be used to isolate the "Archetype Sentinels." These archetype sentinels can then be used to train a perceptron neural network to determine classification of unrelated rats with the same genetic background. This approach can be generalized and used to classify other disease and normal rat models. In addition, the method may be applied to human population studies in a similar manner.

II. BACKGROUND

A. Hypertension

It is estimated that in the U.S., between 50 and 60 million, or about 20%, of people in the United States have hypertension or elevated blood pressure. Known as the 'silent-killer' because of its relative lack of symptoms, persistent hypertension results in distress on the cardiovascular system leading to damage to the eyes, heart, kidneys, or brain, which can ultimately lead to congestive heart failure, heart attack, kidney failure or stroke. While the disease is common, knowledge of the disease's cause is not. Between 90 and 95% of the hypertensive subjects have no specific known cause for the disease. In this case, it is known as primary or essential hypertension. The remaining subjects have secondary hypertension, which is a direct cause of another malady such as disorders of the adrenal glands[4.], kidneys [3.] or arteries [2.]. There is also high correlation between hypertension and smoking, as well as diabetes. In addition, there are other suspected correlations such as genetic factors, stress, obesity[16.], insulin intolerance[5.], hyperinsulinemia[5.], and reduced HDL cholesterol and elevated LDL cholesterol in middle-aged subjects[6.], elevated triglycerides, or left ventricular hypertrophy [7. & 8.]. The research community is exploring this plethora of factors to further access the risk of hypertension relative to a specific individual.

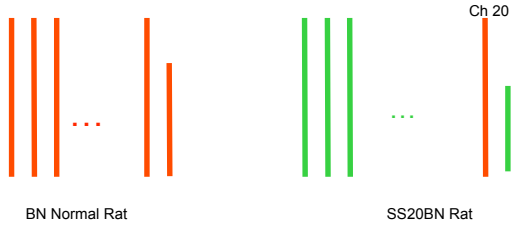

Figure 1 – This shows a pictorial representation of a BN Normal rat (left) and SS20BN (right) Rat, a genetic Salt Sensitive (SS) rat with chromosome 20 replaced with a BN chromosome.

B. Animal Models

The study of hypertension and its underlining causes and affects is very complicated in human subjects. For this reason, animal models are used. Through basic research and genetically controlled experiments, it is possible to look at how measured parameters (phenotypes) differ as a result of genetic variations in animal subjects. One of the common animal models used in genetic study is the laboratory rat. In the last 30 years, more than 500,000 publications used the rat as an experimental species. Rats, mice and humans share most of the basic biochemical pathways of their genomes,[9.,10.,&11.] but extensive research has been placed on characterizing the rat genome [9.,10.,&11.] and on finding homologs within the rat that exist in the human genome.[9.] Further, a number of successful institutions and corporations have developed the ability to create specific genetic configurations.[14.,15.&16.] This allows researchers to use the rat animal model and then relate this to the human physiological system without using human subjects for basic research. Hence, rats are used extensively in scientific experiments in bioinformatics, comparative models for genetic disease, developmental biology, gene structure and function, genetic variability, medical molecular genetics, molecular aspects of infectious disease, pharmacogenetics, proteomics, and tumor biology.[10.,15.&16.]

Through these genetically controlled animal experiments, we can create genetic variations[9.] in subjects and then measure corresponding differences

Renal Database					
Row		SS-20BN Phenotypes	BN Phenotypes	SS20BN Subjects	BN Subjects
1	Original Number	41	41	39	33
2	Sparse Phenotypes	23 (56%)	No Adds		
3	Phenotypes Used	18 (44%)	18 (44%)		
4	Sparse Subjects			9 (23%)	5 (15%)
5	Usable Subjects (M/F)			30 (77%)	28 (85%)
6	Usable Female Subjects			17 (44%)	15 (45%)
7	Max # of NAs	3	1	2	4
8	Archetype Sentinel Subjects			5 (13%)	6 (18%)

Table A – Renal Database Distillation

The first line shows the original number of phenotypes (2nd and 3rd column) and the original number of subjects. The second line shows that 23 SS20BN phenotypes that are considered sparsely populated, and no additional phenotypes are sparsely populated in the BN group. There are nine sparsely populated subjects in the SS20BN group and five sparsely populated BN subjects. Subtracting the number of sparsely populated subjects (row 4) from the original number of subjects (row 1) yields the total number of usable subjects, both male and female. The next row shows the number of usable female subjects. (Subtracting the number of usable female subjects (row 6) from the number of usable subjects both male and female (row 5) yields the number of male subjects that could be used. Row 7 shows the maximum number of NA (not available) values for any given phenotype. It is these NAs that will be filled in with averages for the SS20BN Females or the BN Females in the analysis process. The next row (8) shows the number of archetype sentinel subjects that selected to train the neural network.

that result in the phenotypes. In this study, we are looking at the phenotypes measured relative to normal and hypertensive rats. The normal rat is a Brown Norway (BN) genetic strain, and a hypertensive consomic rat is a salt sensitive (SSN) genetic rat with chromosome 20 from the BN rat replacing chromosome 20 from the SSN rat yielding the SS20BN rat (Figure 1). Phenotypic data for the two genetic strains from the Medical College of Wisconsin (MCW) Rat Genome Database (RGD)[25.] is compared in this study.

III. METHODOLOGY

A. Data Source

The data used is from the PhysGen Data Base sponsored by the Medical College of Wisconsin Phenotypic data is collected for various consomic rats for a number of physiological groupings; lungs, heart, kidneys, etc. We selected the renal and cardiac data sets for the BN (or normal rats) and the SS20BN hypertensive rats. As stated above, it is hypothesized that because these two genetically different rats have different genetic makeup, they will exhibit different phenotype presentations for hypertension as well.[2. & 3.] Additionally, because it is hypothesized that hypertension is a disease caused by a number of factors, but most importantly the kidneys and heart, we will use both the cardiac and renal phenotypes when analyzing these SS20BN hypertensive and BN normal subjects.

B. Nature of Data

Pre-processing of data, particularly with biological subjects, is critical to analysis. Due to the nature of the research process, there are many variables that are not measured or cannot be measured for all subjects and all phenotypes. While elimination of these skews the results for specific phenotypes, so does using incomplete data or substituting incomplete data.

For several reasons, there is incomplete data for various descriptors (phenotypes) or incomplete data for specific subjects. Some phenotypes are of little or no interest to the researcher for certain consomic rat strains. In this case, data is taken sporadically or not at all, resulting in incomplete data (NAs – Not Available) for any given consomic rat strain. Other phenotypes might be particularly hard to acquire. Not all biological readings are easily obtained because of the size and nature of the subject. Others might be considered unimportant as the research progresses, and further accumulation of a specific phenotype might be abandoned, or new phenotypes may be added as research progresses. This leaves older subjects with incomplete data (NAs). Relative to incomplete data in subjects, there is always the chance that the subject expires prior to the completion of the experiment or that other circumstances result in the inaccessibility of the subject for the data accumulation, indicating a problem with the subject. These subjects need to be discarded, since they present an incomplete phenotypic picture or incomplete subject.

Cardiac Database

Row		SS-20BN Phenotypes	BN Phenotypes	SS-20BN Subjects	BN Subjects
1	Original Number of Rats	22	22	29	19
2	Sparse Phenotypes	1	0		
3	Phenotypes Used	21 (95%)	21 (95%)		
4	Sparse Subjects			12 (41%)	5 (26%)
5	Usable Subjects (M/F)			17 (59%)	14 (74%)
6	Female Subjects Used			6 (21%)	7 (37%)
7	Max # of NAs	0	2	0	7
8	Archetype Sentinel Subjects			3 (10%)	4 (21%)

Table B – Cardiac Database Distillation

The first line shows the original number of phenotypes (2nd and 3rd column) and the original number of subjects. The second line shows are 23 SS20BN phenotypes that are considered sparsely populated, and no additional phenotypes are sparsely populated in the BN group. There are nine sparsely populated subjects in the SS20BN group and five sparsely populated BN subjects. Subtracting the number of sparsely populated subjects (row 4) from the original number of subjects (row 1) yields the total number of usable subjects, both male and female. The next row shows the number of usable female subjects. (Subtracting the number of usable female subjects (row 6) from the number of usable subjects both male and female (row 5) yields the number of male subjects that could be used. Row 7 shows the maximum number of NA (not available) values for any given phenotype. It is these NAs that will be filled in with averages for the SS20BN Females or the BN Females in the analysis process. The next row (8) shows the number of archetype sentinel subjects that selected to train the neural network.

C. Pre-Processing of Data

First, we acquire the Cardiac, Renal_A, and Renal_B phenotype data from the PhysGen database at http://pga.mcw.edu. We select 'All Phenotypes'. Then select 'BN 21% O_2 Female' and 'BN 21% O_2 Male' for the BN or normal subjects and 'SS-20BN/Mcwi Female' and 'SS-20BN/Mcwi Male' for the consomic SS20BN hypertensive strain. This provides a total data base (both Female and Male subjects) of all subjects and all phenotypes for both the BN normal and consomic SS20BN hypertensive strains. The PhyGen database is an ongoing research project, so the number of subjects or type and number of phenotypes may vary.

To clean the data, we delete sparsely populated phenotypes, and then delete sparsely populated subjects. In the Cardiac database for female rats, only one phenotype needs to be eliminated, 'rat cardiac body weight.' For this phenotype, all the female SS20BN subjects have no data or NAs.

The Renal database is more complex. We began with 33 BN subjects. Phenotypes with five or more subjects (greater than 15%) with data not available (NA) are eliminated. Eleven phenotypes have more that 20 fields (greater than 60%) not available. The other 12 have NA fields between 15% and 27%. This results in 23 phenotypes being eliminated. (*)

Further, there are differences in the male and female subjects both genetically and as a result, phenotypically as well. It is for this reason that analysis is segregated by gender. This further reduces the pool of subjects available for the characterization process. In this analysis we will be using female subjects. Tables A and B (rows 5 and 6) show the total number of subjects, both male and female, and the number of female subjects, respectively.

D. Ad hoc Method for NA Data

Once these eliminations have occurred and the data is segregated by gender and genetic makeup (BN or SS20BN), 'blanks' or NA (not available) in the remaining data are replaced using an ad hoc method (inserting of the average value) for the gender-specific phenotype remaining. That is, the average for a phenotype is calculated by determining the average for a specific phenotype for female rats of a specific genetic makeup (BN or SS20BN). [18.&19.]

E. Normalization of Data

The procedure outlined above produces a raw data set for each of the subject groups (normal BN rats or hypertensive SS20BN rats). Each phenotype is normalized to produce a value between zero and one. While it is possible to evaluate each phenotype using the raw data, it is imperative that the phenotypes be normalized to categorize them effectively when the subjects are categorized across all phenotypes. With incomplete data being replaced and normalization accomplished, we are ready to categorize or classify each phenotype in a standalone fashion.

F. Determination of 'Archetype Sentinels'

To determine how often each rat is classified or characterized correctly, (i.e. BN rats correctly categorized as BN, and SS20BN rats correctly categorized as SS20BN), we look at each of the phenotypes normal BN rats or hypertensive SS20Bn rats individually using Fuzzy Cluster Means (FCM).

Taking each normalized phenotype, we classify the rats as either normal BN or SS20BN hypertensive based upon a single phenotype. For the 39 phenotypes (18 for cardiac phenotypes and 21 for renal phenotypes), FCM will be used 39 different times, one for each phenotype, to determine whether the rat is classified correctly (BN as BN and SS20BN as SS20BN).

FCM classification results are shown in Tables C and D, for Renal and Cardiac phenotypes respectively, showing the number of correct classifications (signified by an 'x'). After the tables are generated, the number of correct classifications is tabulated for each rat and shown as a percentage in the rightmost column. It is these percentages that are used in the 'selection' process for the Archetype Sentinel subjects.

By selecting the subjects who are characterized correctly most often on a phenotype by phenotype basis (Tables C and D), we are able to isolate Archetype Sentinel subjects in both Renal and Cardiac databases. First we select those subjects which have the highest classification success rates.

For the renal phenotypes, subjects which classify correctly more than 70% of the time where selected as Archetype Sentinels. This yields five renal Archetype Sentinels SS20BN hypertensive subjects and six renal Archetype Sentinels BN normal subjects that are used to train the neural network. Ten renal SS20BN hypertensive subjects and 11 renal BN normal subjects are

---

The first eleven phenotypes have more than 60% of the subjects with NA fields. Phenotypes 12 through 23 have between 15% and 27% NA fields. (1) high salt creatinine clearance (ml/min), (2) low salt creatinine clearance (ml/min), (3) change in heart rate with salt depletion (beats/min), (4) low salt heart rate (beats/min), (5) change in mean arterial pressure with salt depletion (mmHg), (6) low salt mean arterial pressure (one day of recording following salt depletion) (mmHg), (7) high salt plasma creatinine (mg/dL), (8) low salt plasma creatinine (mg/dL), (9) change in plasma renin activity with salt depletion (ls-hs) (ng angl/ml/hr), (10) high salt plasma renin activity (ng angI/ml/hr), (11) low salt plasma renin activity (ng/ml/hr), (12) baseline MAP for NE dose-response relationship (mmHg), (13) delta HR to 0.2 ug/kg/min NE (beats/min), (14) delta HR to 0.5 ug/kg/min NE (beats/min), (15) delta HR to 1.0 ug/kg/min NE (beats/min), (16) delta HR to 0.1 ug/kg/min NE (beats/min), (17) high salt heart rate (beats/min), (18) pre to post control delta HR following NE (beats/min), (19) delta MAP to 0.2 ug/kg/min NE (mmHg), (20) delta MAP to 0.5 ug/kg/min NE (mmHg), (21) delta MAP to 1.0 ug/kg/min NE (mmHg), (22) delta MAP to 0.1 ug/kg/min NE (mmHg), (23) pre to post control delta MAP following NE (mmHg).

	Renal Phenotype																			
Rat Type	baseline MAP for AngII dose-response relationship [beats/min]	baseline MAP for AngII dose-response relationship [mmHg]	delta MAP to 10 ng/kg/min AngII [beats/min]	delta MAP to 25 ng/kg/min AngII [beats/min]	delta MAP to 50 ng/kg/min AngII [beats/min]	delta MAP to 5 ng/kg/min AngII [beats/min]	high salt heart rate [beats/min]	pre to post control delta MAP following AngII [beats/min]	delta MAP to 10 ng/kg/min AngII [mmHg]	delta MAP to 25 ng/kg/min AngII [mmHg]	delta MAP to 50 ng/kg/min AngII [mmHg]	delta MAP to 5 ng/kg/min AngII [mmHg]	high salt kidney weight as percent [%]	pre to post control delta MAP following AngII [mmHg]	high salt urinary excretion of sodium [mEq/day]	low salt urinary excretion of sodium [mEq/day]	high salt urine protein excretion [mg/day]	high salt urine protein excretion [mg/day]	Correct
BN	x	x	x		x	x			x			x	x	x		x	x	x	67%
BN	x	x		x		x	x		x	x	x	x	x	x		x	x	x	78%
BN	x	x	x	x		x		x	x	x	x	x	x	x		x	x	x	83%
BN	x	x	x					x	x	x	x	x	x	x			x	x	67%
BN			x	x		x	x		x	x	x			x	x		x	x	61%
BN	x	x			x	x		x		x	x		x		x	x	x	x	67%
BN		x					x		x	x	x	x	x	x	x	x	x	x	67%
BN	x	x	x						x	x		x	x	x			x	x	56%
BN	x		x		x	x	x		x	x		x	x		x	x		x	67%
BN		x				x			x	x	x	x	x	x		x	x	x	61%
BN	x	x			x	x			x	x	x	x	x	x			x	x	78%
BN		x				x	x				x	x	x	x	x		x	x	56%
BN	x	x	x	x	x			x	x	x	x					x	x	x	72%
BN	x	x				x		x	x	x	x	x	x	x	x	x	x	x	78%
BN	x		x	x	x	x	x						x		x	x	x	x	67%
BN		x		x	x	x			x	x	x	x	x		x	x	x		67%
BN	x	x	x	x	x			x	x	x							x	x	72%
SS20		x		x	x	x	x		x	x	x	x	x		x			x	67%
SS20			x	x	x		x	x					x		x	x	x	x	56%
SS20		x		x	x		x	x	x				x		x		x	x	56%
SS20	x		x	x	x			x	x	x	x	x	x	x	x				67%
SS20		x			x	x	x	x	x		x	x	x	x	x		x	x	67%
SS20		x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	89%
SS20		x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	89%
SS20	x	x	x	x	x		x	x	x				x		x		x	x	67%
SS20		x	x	x	x				x			x	x		x	x	x	x	61%
SS20	x	x	x	x			x	x	x	x		x	x		x	x	x	x	78%
SS20	x	x			x		x	x			x	x				x	x	x	67%
SS20	x	x	x	x	x	x	x	x		x	x			x	x	x	x	x	89%
SS20			x	x	x		x	x		x				x	x		x	x	61%
SS20	x	x	x	x			x	x				x		x			x	x	50%
SS20	x	x	x	x	x	x	x	x				x		x			x	x	72%
	59%	81%	63%	69%	66%	56%	63%	63%	72%	69%	59%	59%	94%	53%	78%	53%	91%	88%	

Table C – This shows the results of the individualized categorization of the Renal Phenotypes via FCM. In each case, the percentage of correct classification (BN rats classified as BN and SS20BN rats classified as SS20BN) by phenotype (columns) and rat subject (rows) is determined. The percentages by rat subject (rows) are used to selection the Archetype Sentinels. The 'archetype sentinels' are highlighted in grey, horizontally. Also shown on the bottom row are the percentages of correct classification (BN rats classified as BN and SS20BN rats classified as SS20BN) by phenotype. The highlighted phenotypes are those that have above an 80% correct classification percentage. These columns are also highlighted in grey.

reserved for creation of the database for testing of the neural network.

We have a limited number of cardiac subjects. In the case of cardiac SS20BN rats, in general we have lower correct classification percentages. To make sure that there are enough subjects for testing the neural network, our selection of the cardiac Archetype Sentinels for the cardiac phenotypes is slightly different. For the cardiac SS20BN subjects, those who are classified correctly more than 67% of the time are chosen, giving us four cardiac Archetype Sentinel SS20BN hypertensive subjects whose phenotype data can be used to train the neural network and two cardiac SS20BN hypertensive subjects that are available for testing purposes.

For the cardiac BN normal subjects, the classification percentages are higher, but the number of available subjects is only seven. Therefore, we have selected three subjects that have correct classifications percentages above 80% and one subject that has a classification percentage at 76%. This gives us 4 subjects as Archetype Sentinels. There is a second rat whose classification is 76%, but this subject was randomly chosen to remain in the test group, simply because the available data pool of subjects is so small. This gives us 4 Archetype Sentinel BN normal subjects whose phenotypic data can be used to train the neural network and three BN normal subjects for testing.

This method shows the actual selection process in this case and the reasons for not adhering to a pure percentage. These are small samples, so tradeoffs

	Cardiac Phenotype																			
Rat Type	baseline peak pressure [mmHg]	baseline time to peak of contraction [sec]	baseline time to peak of contraction [sec]	post-ischemic coronary flow rate [ml/min]	post-ischemic oxygen leakage [IU/g wet wt]	post-ischemic heart rate [beats/min]	post-ischemic infarct size [% LV]	post-ischemic left ventricle developed pressure [mmHg]	post-ischemic left ventricle diastolic pressure [mmHg]	post-ischemic left ventricle systolic pressure [mmHg]	post-ischemic recovery Coronary flow rate [% recovery]	post-ischemic recovery developed pressure [% recovery]	post-ischemic recovery heart rate [% recovery]	post-ischemic recovery systolic pressure [% recovery]	pre-ischemic coronary flow rate [ml/min]	pre-ischemic oxygen leakage [IU/g wet wt]	pre-ischemic heart rate [beats/min]	pre-ischemic left ventricle developed pressure [mmHg]	pre-ischemic left ventricle diastolic pressure [mmHg]	pre-ischemic left ventricle systolic pressure [mmHg]	Correct
BN	x	x	x		x	x	x	x	x		x		x	x	x	x	x	x	x	81%
BN	x	x	x		x	x	x	x		x		x		x	x	x	x		x	76%
BN	x	x	x	x		x	x		x			x	x	x	x	x	x	x		71%
BN	x	x	x		x	x	x			x	x			x	x	x	x	x		67%
BN	x	x	x	x	x	x	x	x	x				x	x	x	x	x	x	x	90%
BN	x	x	x		x		x	x	x	x	x	x	x	x	x	x	x	x	x	90%
BN	x			x			x	x	x	x		x	x	x	x	x	x	x	x	76%
SS20	x	x	x	x	x	x						x				x	x	x		52%
SS20		x	x	x		x				x	x	x	x	x	x		x		x	67%
SS20	x	x	x	x			x				x	x			x	x		x	x	57%
SS20	x	x	x	x	x	x	x	x	x				x	x	x	x		x		81%
SS20	x	x		x		x	x	x			x			x	x	x	x		x	67%
SS20	x	x	x		x	x	x				x			x	x		x	x	x	67%
	92%	92%	85%	62%	62%	77%	77%	62%	62%	54%	54%	54%	69%	69%	85%	92%	69%	100%	69%	85%	54%	

Table D – This shows the results of the individualized categorization of the Cardiac Phenotypes via FCM. In each case, the percentage of correct classification (BN rats classified as BN and SS20BN rats classified as SS20BN) by phenotype (columns) and rat subject (rows) is determined. The percentages by rat subject (rows) are used to selection the Archetype Sentinels. The 'archetype sentinels' are highlighted in grey, horizontally. Also shown on the bottom row are the percentages of correct classification (BN rats classified as BN and SS20BN rats classified as SS20BN) by phenotype. The highlighted phenotypes are those that have above an 80% correct classification percentage. These columns are also highlighted in grey.

must made to select the best Archetype Sentinels used to train the neural network while accommodating the need for creation of the test set.

	Combine Renal and Cardiac 'Archetype Sentinel' Subjects				
		Cardiac Subjects		**Renal Subjects**	
		SS20BN	BN	SS20BN	BN
1	'Archetype Sentinel' Subjects	3	4	5	6
2	Repeated Cardiac Subjects	2	2		
		SS20BN Subjects		**BN Subjects**	
3	Combined 'Archetype Sentinel'	5		6	

Table E – This shows the merging of cardiac and renal archetype sentinels into the combine SS20BN and BN Archetype Sentinel subjects that are used to train the preceptron neural network. We begin with the cardiac and renal subjects shown in row one. By repeating the cardiac phenotype data several times (row 2), we are able to merge the phenotypic data to yield the number of combined archetype sentinels shown in row three.

G. Merging of Renal and Cardiac Phenotypes

Next, we merge Renal and Cardiac databases for both the Archetype Sentinels and the test set. In each case, both the training and the test subjects, there are more renal subjects than there are cardiac subjects. However, because of the genetic breeding of these subjects, we can assume that they are all genetically the same.[17] Therefore, we can combine the cardiac and renal phenotypic data randomly. Since there are more renal subjects than there are cardiac subjects, some of the resulting merged subjects, both 'Archetype Sentinels' and test subjects, will have the same cardiac phenotypic data. While this not ideal, it provides a reasonable number of subjects for both neural network training and testing.

Merging of the renal and cardiac phenotypes to create a combined renal/cardiac Archetype Sentinels or test subject is achieved by concatenating the two sets of values, renal and cardiac, into a single matrix with 39 phenotypic elements. This provides us with five hypertensive SS20BN Archetype Sentinels and six normal BN Archetype Sentinels characterized by 39 phenotypes (18 renal and 21 cardiac). Data from two of the cardiac Archetype Sentinels are repeated because of the disparity between the size of the cardiac and renal databases. (See Table E.)

The same technique is used to merge the test data set. Here the disparity between the number of subjects in the renal and cardiac database is even greater. There are 10 renal SS20BN test subjects and two cardiac SS20BN test subjects. This means the of the data from the two SS20BN cardiac test

subjects are used five times each to yield a total of 10 merged test subjects each possessing 39 phenotypic characteristics. For the 11 renal BN test subjects, there are three cardiac BN test subjects. In this case, phenotypes for two of the cardiac subjects are repeated three times, and one of the cardiac subjects is repeated twice. The resulting test data base is made up of a total of 21 test subjects; 10 merged SS20BN and 11 merged BN all with 39 phenotypic descriptors. (See Table F.)

Combine Renal and Cardiac Test Subjects					
Row		Cardiac Subjects		Renal Subjects	
		SS20BN	BN	SS20BN	BN
1	Test Subjects	2	3	10	11
2	Repeated Cardiac Subjects	8	7		
		SS20BN Subjects		BN Subjects	
3	Test Subjects	10		11	

Table F –Merging cardiac and renal test subjects into the combined SS20BN and BN test subjects used to test the preceptron neural network, we begin with the cardiac and renal subjects shown in row one. By repeating the cardiac phenotype data (row 2), we can merge the phenotypic data to yield the number of combined test subjects shown in row three.

H. Verification of Archetype Sentinels

Using the same FCM program that was used to determine the characterization to determine the Archetype Sentinel subjects (outlined in part '*F. Determination of Archetype Sentinels*'), the Archetype Sentinels are tested to determine how when they were classified. i.e. SS20BN Archetype Sentinel classified as SS20BN subjects or BN Archetype Sentinel as BN subjects. This verification process yielded 100% correct classification of the 11 Archetype Sentinels (6 BN Archetype Sentinels and 5 SS20BN Archetypes Sentinels) selected, each characterized by 39 different phenotypes (21 cardiac and 18 renal).

I. Neural Network Training

Our next step is to use the Archetype Sentinels to train a NN and then test the validity of the NN with the remaining test subjects. A simple perceptron neural network with a hardlim output was generated and trained with the Archetype Sentinels using both the 'adapt' and 'train' modes. (There was no difference in the results based upon the training method used.). Once the network was trained with the Archetype Sentinels, then the 21 test subjects were classified on the trained network. The 21 test subjects with 39 phenotypes each were classified via the simple perceptron hardlim neural network

J. Programs Used

EXCEL was used to manipulate and pre-process the data using functions such as COUNTing NA fields, calculating averages and normalization. MATLAB was for both the FCM program and the NN hardlim perceptron. To ensure consistent program use, the FCM program was designed to accept the filename and location, worksheet designation, and field location for the data being classified. The same program was used throughout the analysis. The only thing changing was the location of the data. The simple NN hardlim perceptron program was written with 39 inputs, one for each of the phenotypes. Training was achieved with 'adapt' and 'train' commands. There was no difference in the result based upon training.

IV. RESULTS

The results of this exercise yield correct classification of 20 of the 21 subjects with 39 phenotypic parameters. That is, all 10 SS20BN subjects were classified correctly as SS20BN, and ten of the 11 BN subjects were classified as BN. One subject was a BN that was classified as a SS20BN. This yields a 95% correct classification, with 5% false positive results.

V. CONCLUSIONS

The biomedical classification of animal models is an important step in the correct interpretation of disease-specific laboratory experiments. Often, in the post-genomic era, large collections of phenotypic data are collected from animal models and human subjects to capture the physiological and pathological basis of human disease. In this work, we have constructed a method that identified archetype sentinel hypertensive and archetype sentinel normal rats, allowing us to train a neural network that could distinguish other subjects with surprising accuracy despite the fact that these subjects were not as prone to correct classification as the archetype sentinel subjects. These results also indicate that with a sparsely populated, multi-dimensional database, it is possible to find Archetype Sentinels, even in small numbers, that can then be used to train a simple perceptron neural network, which then classifies test subjects with a surprisingly high accuracy.

We would also encourage the biological community to see if to these results as an indication of the phenotypes that are most critical in the determination of hypertension in laboratory rats with these two genetic configurations (BN and SS20BN). Further study of biochemical processes and their relationship to hypertension is required.

VI. OTHER APPLICATIONS

While this paper deals with consomic rat bioinformatics data, this Archetype Sentinel technique is not limited to this field. Any complex data set that has both multiple dimensions coupled with small data subject sizes should be considered for this methodology. Possible areas are weather prediction, quality assurance in a manufacturing environment, and failure prediction in both biological and manufacturing environments.

REFERENCES

1. Interleukin-10 Suppresses Tissue Factor Expression in Lipopolysaccharide-Stimulated Macrophages via Inhibition of Egr-1 and a Serum Response Element/MEK-ERK1/2 Pathway, Motohiro Kamimura, Christiane Viedt, Alexander Dalpke, Michael E. Rosenfeld, Nigel Mackman, David M. Cohen, Erwin Blessing, Michael Preusch, Christian M. Weber, Jörg Kreuzer, Hugo A. Katus, Florian Bea, *Circulation Research*. 2005;97:305.
2. Altered Renal Handling of Sodium in Human Hypertension - Short Review of the Evidence, Pasquale Strazzullo; Ferruccio Galletti; Gianvincenzo Barba, *Hypertension*. 2003;41:1000.
3. Conference on Electrolyte and Adrenal Factors in Human and Experimental Renal Hypertension, David F. Bohr, *Circulation*. 1958;17:771.
4. Insulin Resistance and Hypertension - The Insulin Resistance Atherosclerosis Study, Mohammed F. Saad; Marian Rewers; Joseph Selby; George Howard; Sujata Jinagouda; Salwa Fahmi; Dan Zaccaro; Richard N. Bergman; Peter J. Savage; Steven M. Haffner, *Hypertension.* 2004; 43:1324.
5. Glucose-Cholesterol Interaction Magnifies Coronary Heart Disease Risk for Hypertensive Patients, Hillel W. Cohen; Susan M. Hailpern; Michael H. Alderman, *Hypertension*. 2004;43:983.
6. Correlates of Left Atrial Size in Hypertensive Patients With Left Ventricular Hypertrophy - Correlates of Left Atrial Size in Hypertensive Patients With Left Ventricular Hypertrophy, Eva Gerdts; Lasse Oikarinen; Vittorio Palmieri; Jan Erik Otterstad; Kristian Wachtell; Kurt Boman; Björn Dahlöf; Richard B. Devereux, *Hypertension.* 2002; 39:739.
7. Epidemiological Findings Imply That Goals for Drug Treatment of Hypertension Need to Be Revised (Editorial), M. E. Safar, *Circulation.* 2001;103:1188.
8. Exploring a new definition of hypertension. *Rev Cardiovascular Medicine,* M.A. Weber, 2005 Summer;6(3):164-72.
9. Rat Genome Data Base, 2003, Medical College of Wisconsin, Bioinformatics Research Center, http://rgd.mcw.edu/.
10. The Human Genome: Genetic Testing and Animal Models**,** Alexander B. Niculescu, III, M.D., Ph.D., and John R. Kelsoe, M.D**.** *American Journal of Psychiatry* 158:1587, October.
11. The Institute for Genomic Research (TIGR) Rat Gene Index; Sequence Homolog Search, Sequence Reports, and Functional Annotation and Analysis (including GO Annotation), http://www.tigr.org/.
12. Functional Genomics and Rat Models, Jacob, Howard, *Genome Research,* Vol. 9, Issue 11, 1013-1016, November 1999.
13. Charles River Laboratories Inc., Corporate Summary, 2006, http://www.criver.com/research_models_and_services/.
14. PhysioGenix Inc., Corporate Summary, 2006, http://www.physiogenix.com/about.html.
15. Xenogen Inc., Corporate Summary, 2006, http://www.xenogen.com/wt/page/pdf_library.
16. Overweight, Obesity, and Blood Pressure: The Effects of Modest Weight Reduction, Ilse L. Mertens and Luc F. Van Gaal *Obesity Research.* 8:270-278 (2000).
17. PhyGen Data Base, 2003, Medical College of Wisconsin, Bioinformatics Research Center, http://pga.mcw.edu/.
18. Treatment of missing data values in neural network based decision support system for acute abdominal pain, Pesonen E., Eskelinen M., Juhola, M., *Artificial Intelligence Medical*, 1998 Jul;13(3):130-46.
19. Missing Data, Allison, Paul, Sage University Paper, 2002.