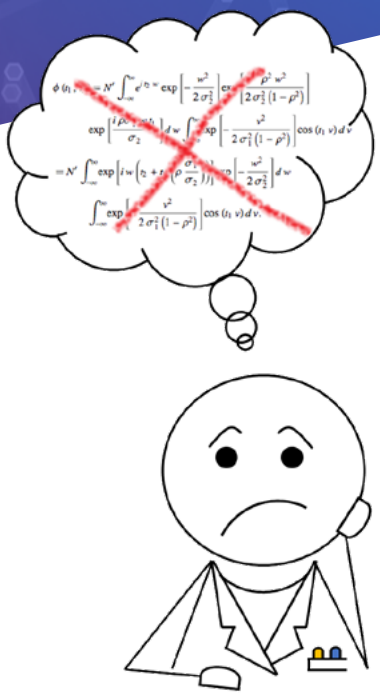
BIOSTATISTICS & BIOINFORMATICS

Common Biostatistical Methods Explained



In the biology field, we often ask: **how does this sample group compare to another sample group?** Biostatisticians and bioinformaticians can help answer this question by applying complex algorithms or concepts. They love the math, but… most of us simply want to understand the basic principles behind these analyses.

In this article, a general understanding of common biostatistical methods is provided.

No math and no advanced degree required!

Hierarchical Clustering

What is it? Hierarchical clustering characterizes how similar (or dissimilar) the samples are based on overall patterns of measurements. For example, the groups may be patients and the overall patterns may be derived from the protein expression across numerous proteins. Hierarchical clustering analyzes the similarity in a binary fashion starting from one sample.

When is it used? This test is performed to stratify samples. You cannot dictate how many clusters are made.

Hierarchical Clustering: Example Questions
- How similar are cell lines X, Y, and Z based on their expression profile?
- How many subsets of breast cancer are there based on the expression profile?
- Is the expression profile of a treated patient more similar to a healthy patient or a diseased patient?

How does it work? Hierarchical clustering uses an algorithm to create a cluster dendogram, which shows how groups cluster with each other (Figure 1). Using the example given in Figure 1, the steps of creating a hierarchical cluster are:

1) The protein expression for each protein across 8 patients is centered and then "scaled" by taking into account the mean and standard deviation, respectively, of the expression values (Figure 2).

2) The Euclidean distance, or the closest distance between two data points based on value intensity (e.g., protein response), is calculated.

3) The two closest data points are clustered together, which are now treated as one data point. The next two closest data points are clustered together, etc. This continues until all data groups are "merged" into one cluster.

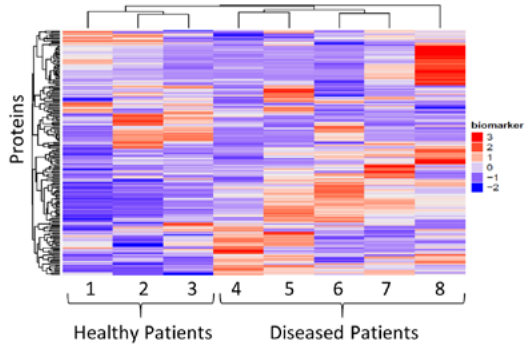


Figure 1. *Hierarchical cluster of healthy and diseased patients where red represents increased expression level and blue represents decreased expression level.*

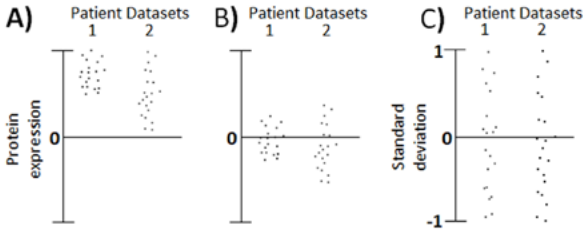


Figure 2. *Example of Centering and Scaling Data. A) Expression level of Protein "X" across two datasets are B) centered and C) scaled so that all datasets have a mean of 0 and a standard deviation of 1.*

What does the data look like? Hierarchical clustering produces 1) a heat map with cluster dendograms (Figure 1) and 2) a table outlining which groups cluster together (table not shown).

K-Clustering

What is it? K-clustering groups samples that are most similar to each other. One cluster (or group) is formed around one centroid; the number of centroids are determined by the user.

When is it used? This test is performed to profile samples. You dictate how many clusters are made.

How does it work?

K-Clustering Example

We analyze the protein profile of 1,000 proteins of 100 breast cancer patients using an antibody-based microarray. We believe that there are five sub-types of breast cancer based on cellular phenotypes. We want to determine whether the patients fall into their diagnosed sub-type.

1) The protein expression for each protein across 8 patients is centered and then "scaled" by taking into account the mean and standard deviation, respectively, of the expression (Figure 3).
2) We tell the software that we want 5 sub-types. The software picks five points on the plot called centroids (Figure 4a).
3) The Euclidean distance, or the closest distance between the centroid and the sample data, is calculated. Numerous iterations are performed to find the optimal centroid position and grouping.
4) The samples are clustered into 5 final groups (Figure 4b). This is the data that we care about.

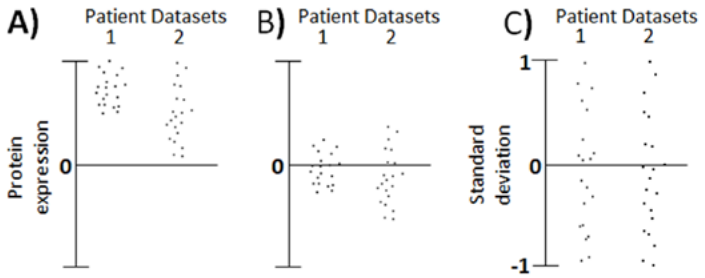


Figure 3. *Example of Centering and Scaling Data. A) Expression level of Protein "X" across two datasets are B) centered and C) scaled so that all datasets have a mean of 0 and a standard deviation of 1.*

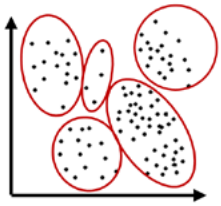


Figure 4a. *2-D plot representing "reduced dimension" data and the clusters assigned in the first iteration. Each spot represents the reduced dimension data from one patient.*

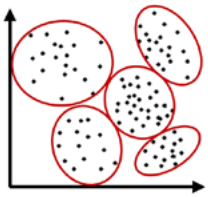


Figure 4b. *2-D plot representing "reduced dimension" data and the clusters assigned in the last (optimal) iteration. Each spot represents the reduced dimension data from one patient.*

What does the data look like? K-Clustering can result in a 2-D plot like Figure 4a and/or as a table that lists the clusters and the samples assigned to each cluster.

Logistic Regression Model

What is it? Logistic regression is a classifier that uses a set of weighted measurements to predict the class (e.g., healthy, diseased) to which a sample belongs based on probability.

When is it used? This model is used when 1) we want to compare > 2 different groups to each other, 2) the samples are independent from each other, 3) the populations may or may not be normally distributed, 4) the population groups are known (e.g., healthy, diseased), and 5) data transformation by other methods results in nonsensical values. It is usually used when the response is binary: yes or no, healthy or diseased, etc.

How does it work?

Logistic Regression Example

We analyze the protein profile of 1,000 proteins of 100 healthy patients and 100 cancer patients using an antibody-based microarray. We want to identify the specific biomarkers that will predict which future patients are healthy or diseased.

1) Center & scale data by subtracting the mean of each patient dataset from itself (Figure 5B) and dividing each patient dataset with its standard deviation (Figure 5C), respectively. Now all datasets have a mean of 0 and a standard deviation of 1.

2) Fit the logistic model based on a subset of variables (Figure 6). This is accomplished by adding different weights to the biomarkers. The data should follow an S-shape (i.e., sigmoid function).

3) Evaluate the performance of the model by receiver operating characteristic (ROC) curve analysis (Figure 7)(see also page 11). Numerous iterations are performed to find the optimal centroid position and grouping.

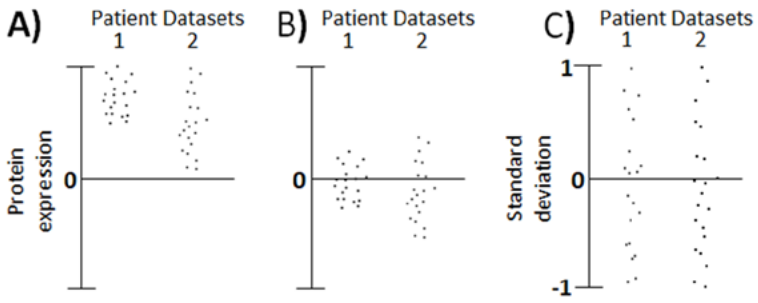


Figure 5. *Example of Centering and Scaling Data. A) Expression level of Protein "X" across two datasets are B) centered and C) scaled so that all datasets have a mean of 0 and a standard deviation of 1.*

Logistic Regression Model (Cont.)

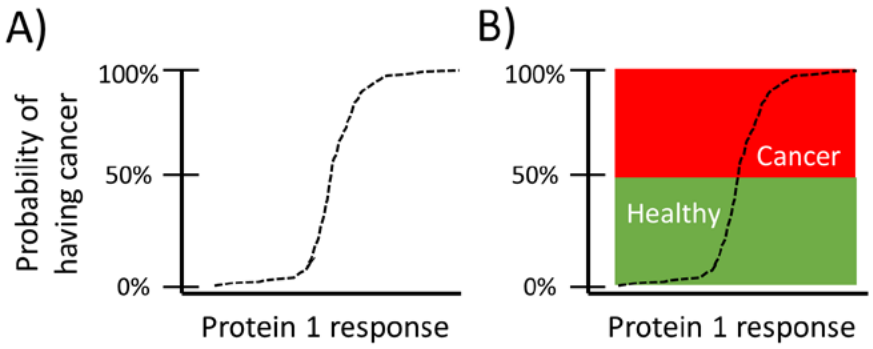


Figure 6. *Protein response across all data is A) plotted and B) assigned to the patient condition (i.e., healthy, cancer). Note that the images in Figures 4 and 5 show 1 protein, but the x-axis may be a combination of proteins.*

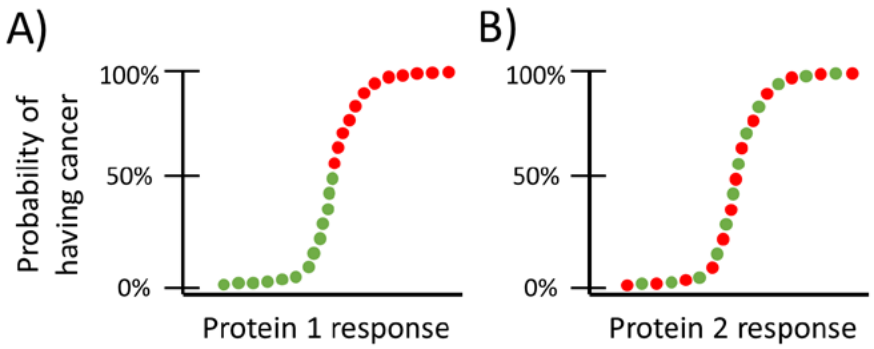


Figure 7. *Examples of biomarkers with different predictive powers to determine the health condition of a patient. A) Protein 1 would be determined to be a good biomarker of cancer while B) Protein 2 would not be a good biomarker. Green = healthy; Red = cancer.*

What does the data look like? The data can be presented in a table format, which would list the biomarkers and their corresponding coefficients (i.e., weights) that are used in the model. Performance of the logistic regression is evaluated by ROC curve analysis.

Random Forest Model

What is it? Random forest consists of hundreds or more decision trees, with each tree using a random subset of data. All of the decision trees cast a vote on the classification of a sample; the majority vote wins.

When is it used? This analysis is one of the most commonly used models. It is used when 1) there are a lot of variables to consider (e.g., expression of thousands of proteins), 2) you only have moderate computing capacities, 3) you don't want to analyze a separate set of samples for cross-validation, and 4) the groups are or are not normally distributed.

How does it work?

Random Forest Example

We analyze the protein profile of 1,000 proteins of 100 healthy patients and 100 cancer patients using an antibody-based microarray. We want to find biomarkers that will predict which future patients are healthy or diseased.

1) Create a table where each row represents a protein and each column represents a patient.
2) Assign groups. Here, you tell the software which samples are healthy and which have cancer.
3) Center data by subtracting the mean of each patient dataset from itself. Now all datasets have a mean of 0.
4) Scale data by dividing each patient dataset with its standard deviation. Now all datasets have a standard deviation of 1.
5) Set aside ⅓ of the data. These samples will be used in Step 7 for cross-validation.
6) Create decision trees using a subset of samples and variables at a time (Figure 8). The modeler determines the number of trees. Each tree will be created using a different random subset of data; the same sample can be chosen more than once to create 1 tree.
7) Determine accuracy of the random forest. The samples set aside in Step 5 are tested against all of the decision trees. The accuracy of the random forest is the proportion of patients that were correctly identified by the random forest.
8) Apply Random Forest to samples with unknown health condition. There will be some trees that classify the patient as healthy, while other trees will classify the patient as diseased. The majority decision wins.

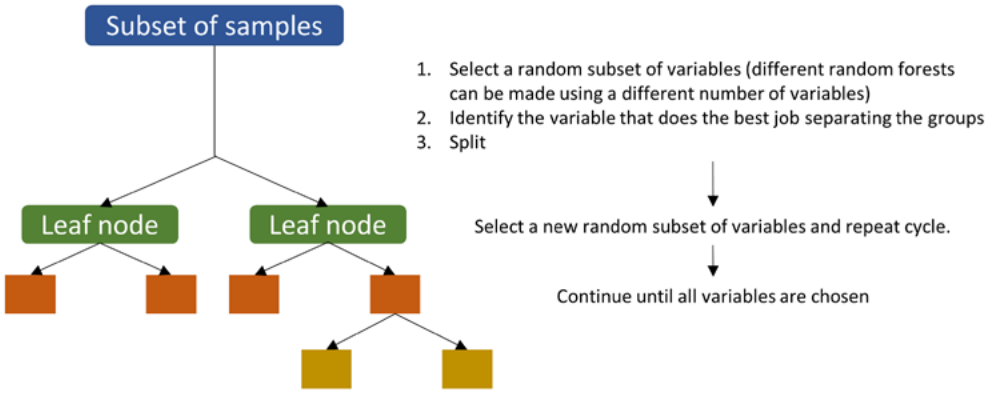


Figure 8. *Creating a Random Forest.*

What does the data look like? The Random Forest model is an ensemble of hundreds of trees that cannot be represented easily; however, the biomarkers used to create the model can be extracted during cross-validation and evaluation of the model's performance (e.g., via ROC curve analysis).

t-test & ANOVA (Analysis of Variance)

What are they? The t-test is a method that determines whether ***two*** populations are statistically different from each other, whereas ANOVA determines whether ***three or more*** populations are statistically different from each other. Both of them look at the difference in means and the spread of the distributions (i.e., variance) across groups; however, the ways that they determine the statistical significance are different.

When are they used? These tests are performed when 1) the samples are independent of each other and 2) have (approximately) normal distributions or when the sample number is high (e.g., > 30 per group). More samples are better, but the tests can be performed with as little as 3 samples per condition.

How do they work?

t-test Example

We want to determine whether the concentration of Proteins 1 – 4 in serum are significantly different between healthy and diseased patients. A t-test is performed, which can be visually explained by plotting the protein concentration on the X-axis and the frequency along the Y-axis of the two proteins on the same graph (Figures 9-12).

Proteins 1 & 2 have the same difference in protein concentration means but different group variances. Alternatively, Proteins 3 & 4 have similar variances but Protein 4 has a larger difference in protein concentration means between the patient groups.

A t-test assigns a "t" test statistic value to each biomarker. A good differential biomarker, represented by little to no overlap of the distributions and a large difference in means, would have a high "t" value.

Which is a better biomarker of disease: Protein 1 or Protein 2? *Protein 1*
Which is a better biomarker of disease: Protein 3 or Protein 4? *Protein 4*

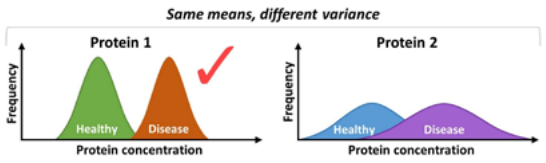


Figure 9. *Overlapping histogram plots for concentrations of protein 1 in different populations.*

Figure 10. *Overlapping histogram plots for concentrations of protein 2 in different populations.*

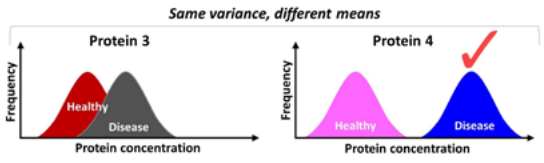


Figure 11. *Overlapping histogram plots for concentrations of protein 3 in different populations.*

Figure 12. *Overlapping histogram plots for concentrations of protein 4 in different populations.*

What type of statistical value do I get? The t-test and ANOVA produce a test statistic value ("t" or "F," respectively), which is converted into a "p-value." A p-value is the probability that the null hypothesis – that both (or all) populations are the same – is true. In other words, a lower p-value reflects a value that is more significantly different across populations. Biomarkers with significant differences between sample populations have p-values < 0.05.

Wilcoxon Rank-Sum

What is it? The Wilcoxon Rank-Sum is a method that determines whether two populations are statistically different from each other based on ranks rather than the original values of the measurements. In other words, it ranks all values to determine whether the values are or are not evenly distributed across both populations.

When is it used? This test is performed when 1) the samples are independent of each other and 2) are or are not normally distributed.

How does it work?

Wilcoxon Rank-Sum Example

We want to determine whether the concentration of Protein 1 in serum is significantly different between healthy and diseased patients. We first collect data for healthy and diseased patients (Figure 13).

1) Wilcoxon Rank-Sum then ranks the values, and assigns the rank to the values (Figure 14)
2) The average ranks from the groups are determined; these averages will be close if there is no difference between the groups.
3) The rank mean of one group is compared to the overall rank mean to determine a test statistic called a z-score. If the groups are evenly distributed, then the z-score will be closer to 0. In this case, the z-score is 3.81, which is equal to a p-value < 0.001. (A p-value of ~0.05 is approximately equal to a z-score of 2.5.)

Is Protein 1 a potential biomarker of disease? Yes, the concentration of Protein 1 in healthy and diseased patients is very different from each other, which is indicated by a very low p-value.

Protein 1 concentration (ng/mL)	
Healthy	Diseased
85.3	69.4
84.3	64.2
79.5	71.4
82.5	71.6
80.2	68.5
84.6	51.9
79.2	72.2
70.9	74.4
78.6	52.8
86.2	58.4
74	65.4
83.7	73.6
Patient # = 12	Patient # = 12
mean = 80.75	mean = 66.15

Figure 13. *Collected protein concentration data for healthy (blue) and diseased (green) patients*

Rank	ng/mL
24	86.2
23	85.3
22	84.6
21	84.3
20	83.7
19	82.5
18	80.2
17	79.5
16	79.2
15	78.6
14	74.4
13	74
12	73.6
11	72.2
10	71.6
9	71.4
8	70.9
7	69.4
6	68.5
5	65.4
4	64.2
3	58.4
2	52.8
1	51.9

Figure 14. *Figure 14. Ranked protein concentrations from high to low collected for healthy (blue) and diseased (green) patients. The average concentration of Protein 1 in the healthy patient is clearly higher than the average concentration of Protein 1 in the diseased patient.*

What does the data look like? Wilcoxon Rank-Sum produces a test statistic value (i.e., z-score), which is converted into a "p-value." A p-value is the probability that the null hypothesis – that both populations are the same – is true. In other words, a lower p-value reflects a value that is more significantly different across populations. Biomarkers with significant differences between sample populations have p-values < 0.05.

Linear Discriminant Analysis (LDA) Model

What is it? Linear discriminant analysis (LDA) separates samples into > 2 classes based on the distance between class means and variance within each class. LDA can also serve to reduce data dimension.

When is it used? This analysis is used when there are a lot of variables to consider (e.g., expression of thousands of proteins). LDA makes a lot of assumptions, such as the 1) sample measurements are independent from each other, 2) distributions are normal, and 3) co-variance of the measurements are identical across different classes. Therefore, LDA will not be accurate if the data do not follow these criteria. Unlike LDA, the support vector machine (SVM) model does not assume anything about data distribution (see page 13).

How does it work?

LDA Analysis Example

We analyze the protein profile of 1,000 proteins of 100 healthy patients and 100 cancer patients using an antibody-based microarray. We want to find biomarkers that will predict which future patients are healthy or diseased. This represents high-dimensional data since each sample is characterized by 1,000 variables. Put another way, the sample point is located in a 1,000-dimension space. We want to find the linear discriminant function that will classify the patients as healthy or diseased.

1) Assign samples to patient groups.
2) Center & scale data by subtracting the mean of each patient dataset from itself and dividing each patient dataset with its standard deviation, respectively. Now all datasets have a mean of 0 and a standard deviation of 1.
3) Create Discriminant Functions (i.e., "Data Reduction").
- **a) Determine # of Discriminant Functions (DFs).** The number of DFs is determined by the number of comparisons between classes (or responses). In this case, there are only 2 classes (i.e., healthy and diseased), so one DF is sufficient.
- **b) Create DFs.** Here, the DF is a function describing the linear boundary that best separates the classes based on the class means and variance within each class (Figure 15A-15B). In other words, the DF is an equation in which the variables (i.e., biomarkers) are weighted differently.

4) Visualize DF (Figure 15C).

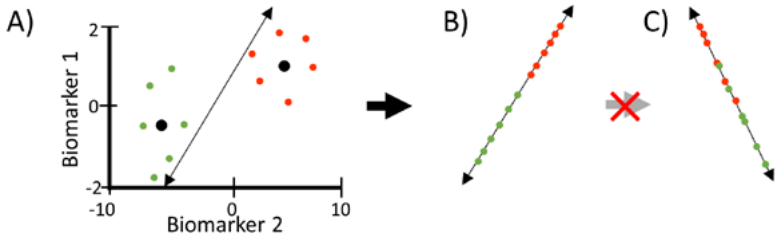


Figure 15. *LDA modeling A) identifies the boundaries between different classes [by taking into account the distance between the class means (black dots) and the variance within each class] to B) create a DF, which is represented by the line. In this example, 2-dimensional data has now been reduced to 1 dimension. C) This DF would not be made because it does not separate the different classes from each other well.*

What does the data look like? LDA analyses are represented as a table listing the weights of biomarkers per DF. LDA performance can be ascertained using ROC curve analysis.

PCA Analysis

What is it? Principle Component Analysis (PCA) transforms high-dimensional data into a lower-dimensional structure to improve data presentation, pattern recognition, and analysis. PCA determines which dimensions will result in the largest variability of measurements (e.g., expression of specific proteins) across all samples. It does not separate the different groups from each other as the method is unsupervised.

When is it used? This analysis is used when 1) there are a lot of variables to consider (e.g., expression of thousands of proteins), and 2) you want to ensure that the transformed measurements are independent of each other. PCA can help discover relationships between biomarkers that may not be intuitive.

How does it work?

PCA Analysis Example

We analyze the protein profile of 1,000 proteins of 4 healthy patients and 4 cancer patients using an antibody-based microarray. This represents high-dimensional data since each sample is characterized by 1,000 variables (or biomarkers). Put another way, the sample point is located in a 1,000 dimension space. We want to find the biomarkers that are the most variable across the samples. We hope that, by doing so, we will be able to identify patterns (i.e., biomarkers that are expressed differently in healthy and diseased patients).

1) **Center & scale data** by subtracting the mean of each patient dataset from itself and dividing each patient dataset with its standard deviation, respectively. Now all datasets have a mean of 0 and a standard deviation of 1.
2) **Create Principal Components (i.e., "Data Reduction").**
   - **a) Determine # of Principal Components (PCs).** The number of PCs is determined by the minimum number of samples or variables. In this case, the number of patients is lower than the number of proteins analyzed, so 8 PCs are made.
   - **b) Create PCs.** Here, each sample has 8 PCs, where each PC value is a weighted summation of the 1,000 variables.
   - **c) Weigh variables differently.** Variables that have the largest variance across patient groups are weighted more than variables with smaller variances. Since the variables are weighted differently for each PC, the PCs are uncorrelated to each other. Now the sample point is placed in an 8-dimensional space. The PCs are ordered based on sample variability, such that the first PC (PC1) has the highest variability, PC2 has the second highest variability, and so on.
   - **d) Reduce data dimension.** We plot the PCs against each other, and we can now use these plots to create new axes via data transformation (Figure 16). Thus, we reduce the space from 1,000 dimensions to 8 dimensions.
3) **Compare PCs visually** using a scatter plot to observe patterns (Figure 17). PC1 and PC2 are usually the PCs that are plotted against each other because they result in the highest and second highest distribution variances, respectively, of all PCs. Groups that are more similar to each other will have higher overlap on the PCA plot. Therefore, group similarity or dissimilarity can be easily and visually discerned using a PCA plot.
4) **PCA analysis.** Biomarkers making greater contributions to the PCs can be identified as they will have higher weights.

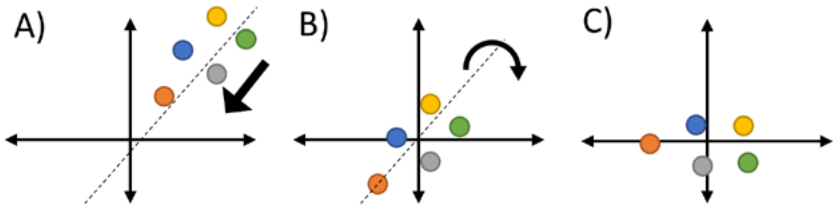


Figure 16. *An example of centering and transforming PC data. A) Scatter plot of one PC versus another PC. Data is B) centered and C) transformed.*

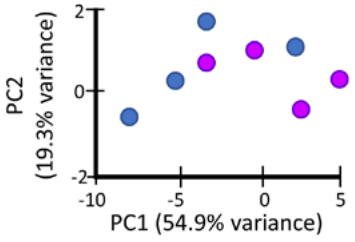


Figure 17. *PCA plot of 1,000 biomarkers among 4 healthy (blue) and diseased (pink) patients. The weights applied to the variables in PC1 result in a wider distribution (i.e., variance) of data points than PC2.*

What does the data look like? PCA analyses are represented as 1) a figure (like Figure 17) or 2) a table listing the first few PCs. Biomarkers that were weighted the most can be considered as potential biomarkers for follow-up validation studies. Since biology is complicated and we still have a lot to learn, PCA analysis may not identify variables that are intuitive.

ROC Curve Analysis

What is it? A Receiver's Operating Characteristic (ROC) curve plots every value of a continuous measurement by its specificity and sensitivity to distinguish health status in a population of subjects. The area under the curve (AUC) reflects the measurement's potential to be a diagnosis tool. At a specific sensitivity, the specificity can be determined, or vice versa.

When is it used? This analysis is used to identify the appropriate classifying thresholds to diagnose a patient with expected sensitivity and specificity.

How does it work?

ROC Curve Analysis Example

We've identified a potential biomarker, Protein "A," of Alzheimer's disease that is elevated in the plasma of Alzheimer's patients compared to healthy patients (Figure 18). We now need to determine the lower cut-off value of "Protein A" levels that will identify a patient with Alzheimer's disease. We don't want to have the threshold too low like concentration X in Figure 18 or else a lot of healthy patients will be wrongly diagnosed. However, we also don't want to have a threshold that is really high like concentration Y in Figure 18 because a lot of Alzheimer's patients won't get diagnosed. A ROC curve can help identify the "sweet spot" (i.e., optimum sensitivity-specificity balance).

Figure 19 explains what sensitivity and specificity are. Ideally, the sensitivity and specificity would be 100%. In reality, virtually all biomarkers do not have perfect sensitivity and specificity.

A ROC curve is generated across all values and the AUC is determined (Figure 20). Higher AUC values represent a better biomarker. A point along the ROC curve is chosen with the desired trade-off between sensitivity and specificity. With known sensitivity and specificity, the cut-off value can be ascertained.

For this example, let's assume the black dashed line is the ROC curve for our data. We would likely choose the X-Y coordinate of 0.1, 0.8, such that the biomarker would have a specificity of 90% and a sensitivity of 80%.

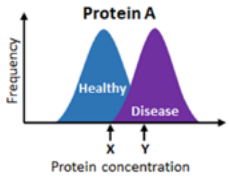


Figure 18. *Overlapping histogram plots for concentrations of Protein A in different populations. A cut-off of concentration "X" will have high sensitivity, but low specificity. A cut-off of concentration "Y" will have low sensitivity, but high specificity.*

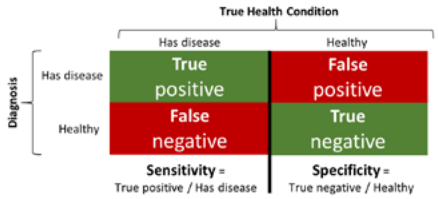


Figure 19. *Calculation of sensitivity and specificity*

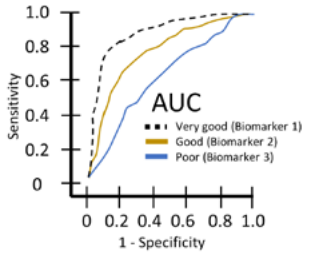


Figure 20. *Comparison of ROC curves across three potential biomarkers. The higher the AUC value, the higher predictive value of the biomarker. Biomarker 3 has very poor predictive power (AUC ~0.5) as it cannot differentiate between healthy and diseased patients at all.*

What does the data look like? ROC curve analyses are usually portrayed as a plot like Figure 20.

SAM (Significance Analysis of Microarray)

What is it? SAM is a method used for large-scale gene or protein expression data like those collected with microarrays. It addresses the issue of analyzing large-scale data in which a microarray experiment of 10,000 proteins would identify 100 proteins by chance using a p-value cut-off of 0.01. Therefore, SAM applies a t-test at the ***individual*** gene or protein level to determine whether the expression pattern for that gene or protein is significant.

When is it used? This test is performed when the samples 1) may not be independent of each other and 2) are or are not normally distributed. It can help identify expression patterns that have little difference between the control and test groups but are nevertheless significant.

How does it work?

SAM Example

We want to find serological proteins that are different between 12 healthy and 12 diseased patients using an antibody-based microarray targeting 1,000 proteins.

1) **The observed relative difference per protein across groups is determined,** which considers the mean and variance of each group (Figure 21). This step accounts for protein-specific fluctuations.
2) **The expected relative difference per protein across groups is determined** by averaging the protein responses across numerous permutations. An example of a permutation is given in Figure 22 in which a group label (e.g., healthy, diseased) is assigned at random. These random permutations form a simulated distribution of expected relative differences (like a t-statistic). The random permutations are also used to calculate the false discovery (FDR), or the rate at which a protein will be incorrectly identified as significant.
3) **Plot the observed vs expected relative difference** (Figure 23). This is a visual way of looking at the data.
4) **Identify proteins-of-interest** that deviate from the diagonal line using a threshold (dashed lines in Figure 23). The threshold is determined by calculating false discovery rates (FDRs) using data from the permutations.
5) **Determine the statistical significance of the proteins-of-interest.** Biomarkers with larger deviations between the observed (step 1) and expected (step 2) relative difference are deemed significant. In other words, the larger the deviation and lower the FDR, the higher the significance.

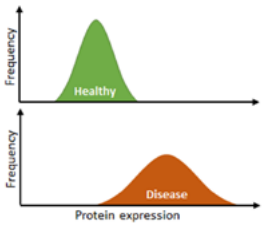


Figure 21. *Histogram plots of Protein 1 expression in different populations.*

Protein 1 Response	
Healthy	Diseased
85.3	69.4
84.3	64.2
79.5	71.4
82.5	71.6
80.2	68.5
84.6	51.9
79.2	72.2
70.9	74.4
78.6	52.8
86.2	58.4
74	65.4
83.7	73.6

permutation

Protein 1 Response	
"Healthy"	"Diseased"
85.3	69.4
64.2	84.3
71.4	79.5
82.5	71.6
80.2	68.5
51.9	84.6
79.2	72.2
74.4	70.9
78.6	52.8
86.2	58.4
65.4	74
73.6	83.7

Figure 22. *Permutation example for Protein 1. Note that an equal number of datasets from healthy (blue) and diseased (green) patients are being compared to each other. The "healthy" and "diseased" data sets would be compared in this permutation. Numerous permutations would be performed.*

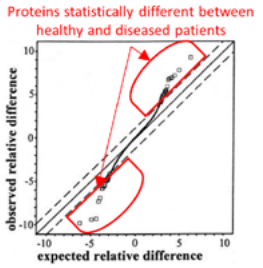


Figure 23. *Scatter plot of observed vs expected relative differences (t-statistic) of a protein. Dashed lines = threshold cut-off. Figure altered from Tusher et al. Proc Natl Acad Sci. 2001 Apr 24; 98(9): 5116-5121.*

What does the data look like? For each gene or protein, SAM produces a test statistic value based upon the observed value's deviation from the expected value. Unlike other models that use a p-value or FDR, SAM determines significance based on the deviation of the observed data from the expected value; the expected value is based on numerous permutations of the original data.

Support Vector Machine (SVM) Model

What is it? Support vector machine (SVM) determines the boundaries that best classifies the different groups from each other using a subset of variables (e.g., biomarkers) in multi-dimensional space. The boundary is a hyperplane in which a subset of data points closest to the hyperplane (called support vectors) have the maximum distance from each other.

When is it used? This analysis is used when 1) there are multiple variables to consider (e.g., expression of thousands of proteins), and 2) regular linear transformations are not enough. Unlike LDA, SVM does not assume anything about data distribution.

How does it work?

SVM Example

We analyze the protein profile of 1,000 proteins of 100 healthy patients and 100 cancer patients using an antibody-based microarray. We want to find biomarkers that will predict which future patients are healthy or diseased.

1) **Create a table** where each row represents a protein and each column represents a patient.
2) **Assign groups.** Here, you tell the software which samples are healthy and which have cancer.
3) **Center & scale data** by subtracting the mean of each patient dataset from itself (Figure 24B) and dividing each patient dataset with its standard deviation (Figure 24C), respectively. Now all datasets have a mean of 0 and a standard deviation of 1.
4) **Fit the SVM model.** This transformation finds the optimal boundary between the groups in multi-dimensional space using a subset of data points, such that the data points from different groups closest to the boundary have the maximum distance from each other (Figure 25). The appropriate number of hyperplanes is determined by cross-validation of SVM model performance.

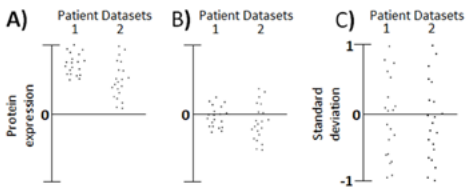


Figure 24. *Example of Centering and Scaling Data. A) Expression level of Protein "X" across two datasets are B) centered and C) scaled so that all datasets have a mean of 0 and a standard deviation of 1.*

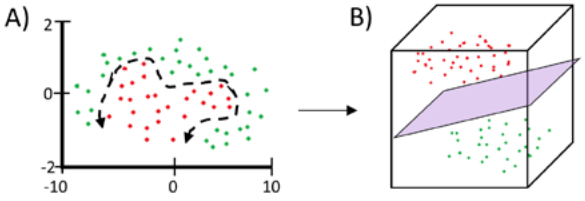


Figure 25. *SVM modeling A) identifies the boundaries between groups and then B) performs complex transformations to find the hyperplane boundary between the groups in multi-dimensional space. In this example, a hyperplane was determined in 3 dimensions.*

What does the data look like? The performance of the SVM model is evaluated using ROC curve analysis. Final SVM model results are represented as a table listing the selected biomarkers that classify the groups from each other.

References

Ashburner M. Gene Ontology: Tool for the Unification of Biology. Nature Genetics. 2000; 25: 25–29. doi: 10.1038/75556

Kanehisa M, et al. KEGG as a Reference Resource for Gene and Protein Annotation. Nucleic Acids Research. 2016; 44 (D1): D457–D462. doi: 10.1093/nar/gkv1070

Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008; 28(5). doi: 10.18637/jss.v028.i05

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. (2018) Available online: https://www.R-project.org/.

Robin X, et al. pROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves. BMC Bioinformatics. 2011; 12:77. doi: 10.1186/1471-2105-12-77

Schwender H. siggenes: Multiple Testing using SAM and Efron's Empirical Bayes Approaches. 2019. R package version 1.60.0.

Szklarczyk D, et al. The String Database in 2017: Quality-Controlled Protein-Protein Association Networks, Made Broadly Accessible. Nucleic Acids Research. 2017;45(D1):D362-D368. doi: 10.1093/nar/gkw937

Tusher GV, et al. Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. Proceedings of the National Academy of Sciences. 2001; 98 (9): 5116–21. doi: 10.1073/pnas.091062498

Yu G, et al. ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A Journal of Integrative Biology. 2012; 16 (5): 284–87. doi: 10.1089/omi.2011.0118