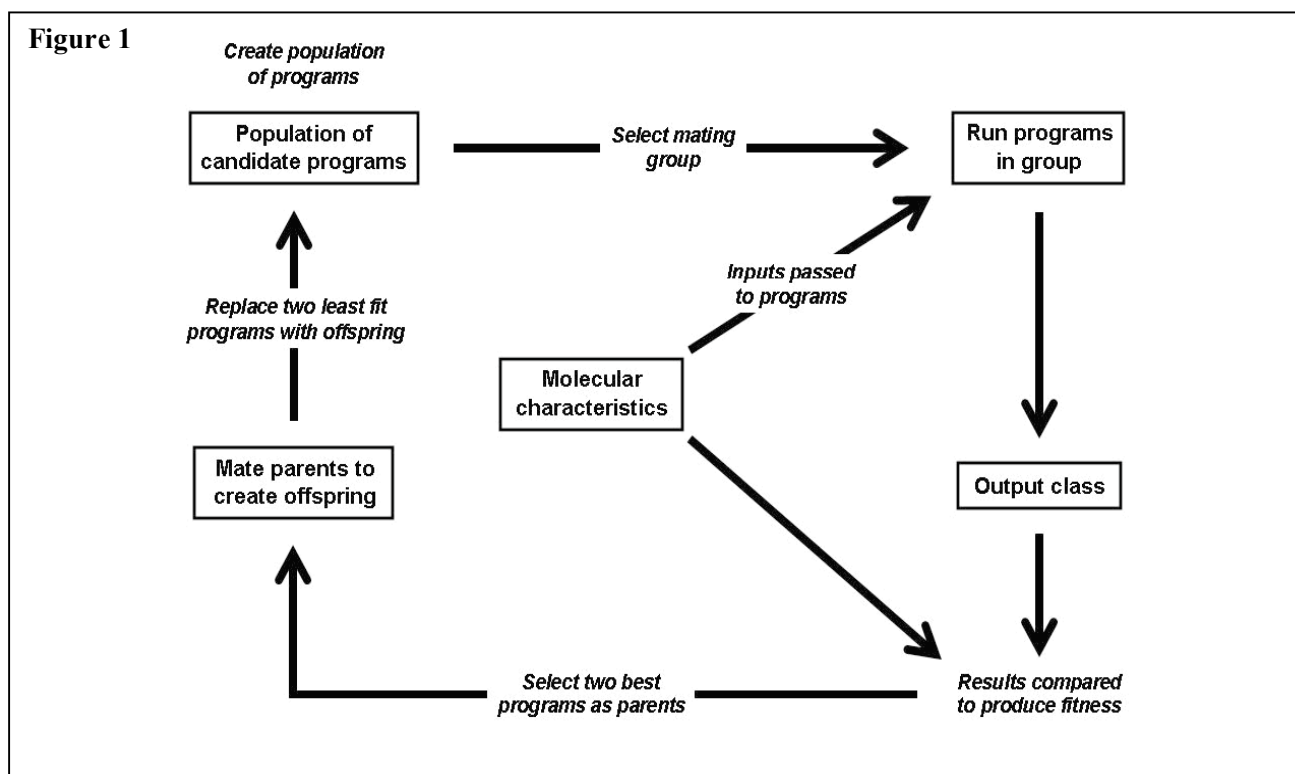


Genetics Squared's *Evolver*™ Technology

Genetic Squared's proprietary analytic technology, *Evolver*, is an implementation of genetic programming (GP) software to the analysis of data in the life sciences. *Evolver* produces mathematical models that describe a dataset based on what specific questions are being asked about that dataset. These models can select variables from any combination of data types, such as protein chemistry, genomic, or demographics. Despite its name, GP, a machine-learning technique, is not based on genetics. Instead, the name derives from the fact that it uses ideas from natural selection and population dynamics. It literally "evolves" a solution to a problem. **Figure 1** shows a high-level flowchart of how the technology works. In it, we use as an example a set of molecular data derived from a known group of good-prognosis patients and a similarly sized group of poor-prognosis patients. The system will learn to differentiate between these two groups by generating and then refining classification rules programmatically.



The Process

The process starts in the upper left hand corner of the diagram where we create an initial population of computer programs that use randomly selected inputs and mathematical operators to construct a programmatic embodiment of a rule that predicts membership in a target class, e.g., patients with good prognosis. Moving to the right, we select a small "breeding population" of programs from the main population and apply these to the molecular data available. Moving downwards, each of these rules makes a prediction about the samples with which it is presented and we compare these predictions with the known truth. This is used to create a fitness measure for each program. Moving left along the bottom of the flowchart, the two best programs are selected and combined to produce offspring programs which, moving back to our starting point, replace the worst programs in the main population. This process is then repeated by going around this loop until some completion criterion is reached. The strategy may be compared to horse breeding where one might take a horse that can run fast and breed it with a horse that has great endurance in hopes of producing

an offspring that can run fast and has great endurance. It is a surprisingly effective technique that has been used in many applications, including modeling chemical processes and kinetics, designing antennas and electronic circuits, and predicting the movement of financial markets.

Advantages of Genetics Squared's Analytic Approach

Evolver produces rules from whatever inputs are provided, which can include any combination of diverse data such as patient demographics, clinical chemistry, histology genomic and proteomic data. This unique ability to combine different types of data allows a more complete description of a patient's baseline status and increases the probability of finding the most important factors that indicate sensitivity to a given drug or good/bad prognoses. Even if molecular data are not available or cannot be obtained there is still a reasonable chance that we can develop a patient profile based on standard clinical data that correlates with a particular clinical outcome and develop a commercially useful therapeutic diagnostic.

Microarrays can have more than 40,000 probes. This presents a unique problem for most approaches, as they often do not scale well to large numbers of inputs. It is, therefore, usually necessary to find some way to reduce the number of inputs before applying techniques such as statistical or cluster analysis. Typically, inputs are reduced by making some assumptions about the data or the target population such as assuming that important genes vary in expression level by a certain amount or that genes associated with particular pathways are important and should be selected for analysis. While these assumptions may well be true, they introduce bias into the selection process, which may result in missing other significant factors. Conversely, the system also allows for preferential selection of variables, if desired. There are often good reasons to bias the system to select variables that are particularly well characterized and may provide results that are more rapidly integrated into the clinic. For example, when attempting to identify diagnostic predictors, it may be advantageous to preferentially select proteins that are known to be secreted into the blood.

The problem of working with large numbers of inputs will only increase as new techniques are developed. Genomic chips capable of detecting more than 500,000 SNPs are an example of how this problem has already outgrown most analytical technologies, particularly as there is no way to select which SNPs are more significant than others because they are simply present or not and there is no fold-variation or other factors to indicate their importance. The **Evolver** approach can be used even in these extreme cases to identify the key factors or combinations of factors in predictive rules. Commercially, this means that we can find diagnostic rules where other approaches are incapable. We have successfully taken data sets with large numbers of inputs and found rules that use fewer than 6 inputs to classify a patient sample. The ability to reduce the number of genes needed for an accurate predictor can decrease the cost of a diagnostic and allow more quantitative technologies, such as RT-PCR, to be used where each additional measurement adds to the cost of the product.

Evolver results are human-readable. Many other machine-learning approaches may be described as black-box solutions where a prediction is made but it is not clear to the user how it was made. Because **Evolver** produces simple rules, the results can be evaluated for biological and chemical relevance and for discovery of novel relationships. These rules can provide insight and generate new, testable hypotheses for exploration.

Evolver is adept at uncovering non-linear relationships among variables. It is well accepted that many factors comprising biological systems interact in ways that are not linear. In order to better understand how these systems work, it is critical to accurately characterize these relationships. Most other analytic platforms have difficulty producing non-linear models, particularly from a large set of inputs.

Genetics Squared's **Evolver** technology:

1. Is not encumbered by preconceptions that limit human problem solving.
2. Integrates diverse, multiple data types, such as demographic and clinical data with specific biochemical markers (i.e. genomic, proteomic, etc.) into a single result. The analysis can also handle missing values in the data.
3. Selects globally from thousands of variables without bias encouraging the discovery of novel factors in the results.
4. Allows for preferential selection of variables and can limit the complexity of the predictive rule by identifying inputs that are most dominant in defining outcome.
5. Uncovers typical, biological non-linear relationships among variables, such as genes or proteins.
6. Provides results in the form of simple rules, utilizing very few variables, which can be evaluated for biological and chemical significance and for discovery of novel relationships.

	Understandable Simple, Readable Results	Unbiased Automatic Selection of	Encompassing Automatic Integration of	Ideally Suited to Biological Systems Non-Linear
--	--	--	--	---

		Variables	Different Data Types	Relationship Characterization
<i>Evolver</i>	Yes	Yes	Yes	Yes
Statistical Analysis	Yes	Limited		Limited
Cluster Analysis	Yes			
Support Vector Machine				Yes
Neural Networks				Yes

Evolver has been used in prognostic development, clinical oncology trial patient stratification and other areas of clinical and preclinical biomedical research such as infectious disease, nephrology, neurology, and autoimmune disease.

An Example

As an example, Genetics Squared collaborated with Dr. Richard Cote’s lab at the University of Southern California to predict nodal involvement in bladder cancer from primary tumor tissue. Seventy genes were selected based on their relevance to bladder cancer. The study included primary bladder tumor tissues from 60 patients across different stages and 5 control tissues of normal urothelium, the latter serving as negative controls. All tissues were profiled using a quantitative RT-PCR technique. The data was divided into a 35-sample training set and a 30-sample validation set.

Genetics Squared produced a rule that was 90% accurate and had a positive-predictive value of 100% for identifying node-positive patients on the validation set. This is typical of Genetics Squared’s ***Evolver***TM technology. ***Evolver***TM produces highly accurate predictive rules, which describes a quantitative relationship among the inputs.

For more details refer to:

Mitra AP, Almal AA, George B, Fry DW, Lenehan PF, Pagliarulo V, Cote RJ, Datar RH, Worzel WP. Use of genetic programming in the analysis of quantitative gene expression profiles for identification of nodal status in bladder cancer. *BMC Cancer* 2006, 6:159.