Case Study

Exagen Diagnostics: Genomic Marker Discovery on an Apple Xserve Computational Cluster

By Frank Lacombe

Executive Summary

Exagen Diagnostics, Inc. uses computational technology to discover genomic markers that form the basis of prognostic and predictive tests used in patient care and pharmaceutical clinical trials.

At the core of Exagen's business is the Exagen Discovery Engine, a proprietary software process running on a high-performance computational cluster. When selecting an Apple dual processor Xserve G5 for this mission-critical role, Exagen placed high value on the ease of use of the Mac OS X operating system, the superior throughput-to-size ratio of the Xserve, and the very low noise level of the cluster in its Xtreme Xrack soundproof rack enclosure. The unsurpassed cross-platform interoperability of Mac OS X Server was also important in a heterogeneous computational environment that includes Mac, Windows, and Linux desktop and notebook computers.

This paper describes the scientific context for the Exagen process, explains the computational approach taken by Exagen, and draws general conclusions that scientists and engineers working in related fields might find applicable to their own problem domains.

Contents

Overview

About Exagen Diagnostics
Exagen Diagnostics is a leader in an emerging class of genomic marker tests that will predict disease progression or response to pharmaceutical drug treatment.

Established in 2002 with Dr. Waneta Tuttle, a leading biomedical entrepreneur, as founding CEO, Exagen is a spin-off of Quasar International, whose technology screens large volumes of data to identify defects in automotive parts. Exagen researchers have adapted and advanced this technology to analyze huge volumes of genomic data, resulting in a proprietary platform that enables the company to dramatically accelerate discovery of marker combinations and create a robust pipeline of products. Exagen discovers, validates, and commercializes genomic markers in multiple therapeutic areas.

Exagen applies its proprietary discovery technology to rapidly identify practical and accurate combinations of genomic markers for testing, ultimately enabling physicians and patients to make better treatment decisions. By partnering with pharmaceutical companies, Exagen can rapidly discover and advance optimal combinations of markers in support of pharmaceutical R&D/Phase IV, identifying significant marker sets in weeks rather than months.

The mission of Exagen Diagnostics, Inc. is to discover and commercialize small sets of genomic biomarkers for use in predictive or prognostic testing applications. Each Exagen solution consists of a small set of biomarkers (typically three to five) that correlates closely to a specific biological condition or medical outcome. The Exagen process has already yielded notable results:

- The breast cancer prognostic marker assays developed by Exagen provide the first DNA-based tests for hormone receptor positive (HR+) and hormone receptor negative (HR-) patients. These tests identify patients with a high or low risk of tumor recurrence, so that patients who are at low risk may choose not to be treated and those who will benefit are treated appropriately.
- Two predictive tests currently under development by Exagen will help the four million people in the United States and almost 200 million worldwide who are infected with the hepatitis C virus (HCV). The first test identifies those patients most likely to respond to the standard treatment regimen; the other identifies which HCV patients show evidence of liver damage.

Pharmaceutical companies can use Exagen tests as part of their drug development process. Physicians can use Exagen tests to identify patients in need of treatment and to help determine which therapeutic measures are appropriate for individual patients. Depending on the application, one test can serve both the medical and pharmaceutical communities or specialized tests can be developed for each.

One common way for pharmaceutical companies to use such tests is for patient stratification. By identifying and eliminating those patients who are unlikely to respond to the drug being tested in Phase IV clinical trials, pharmaceutical companies can enrich the clinical trials for those patients who will benefit most from the treatment. In addition, pharmaceutical companies are forming partnerships with companies such as Exagen to look at biomarkers for application in the drug discovery process and in clinical trials, whether it's in preclinical testing or in Phase II, III, or IV trials. Researchers hope that strong correlations between biomarkers and specific diseases will help identify disease pathways that provide promising targets for new pharmaceutical therapies.

Individual Markers versus Marker Combinations

For years, molecular biologists focused on understanding diseases in terms of single genes. This focus was especially true in the study of cancer. An example of this single-gene focus involved the gene P53, which is related to growth control. The discovery that P53 played a role in a particular kind of cancer led to a flurry of research to see whether other individual genes could play a similar role in other cancer types.

Over the course of several years, however, researchers realized that most complex diseases, including cancer, involve equally complex causes: multiple disease pathways that can involve seemingly unrelated genes and proteins. While the approach of assuming multigene mechanisms for complex disease is more biologically sound, it is also often harder to clearly assign individual functions to the genes identified. Moreover, the mathematical scale of computational problems increases dramatically when, instead of looking for one gene out of a pool of 30,000 or so, researchers must consider combinations of many genes.

This problem of mathematical scale would probably be insurmountable except for the fact that computational power continues to increase rapidly in relation to the real cost of computers. Faster hardware and refinements in analytical processes, neural networks, and data-mining technology now make it possible for innovative researchers to solve in days, and at relatively low cost, computational problems that were infeasible only a few years ago.

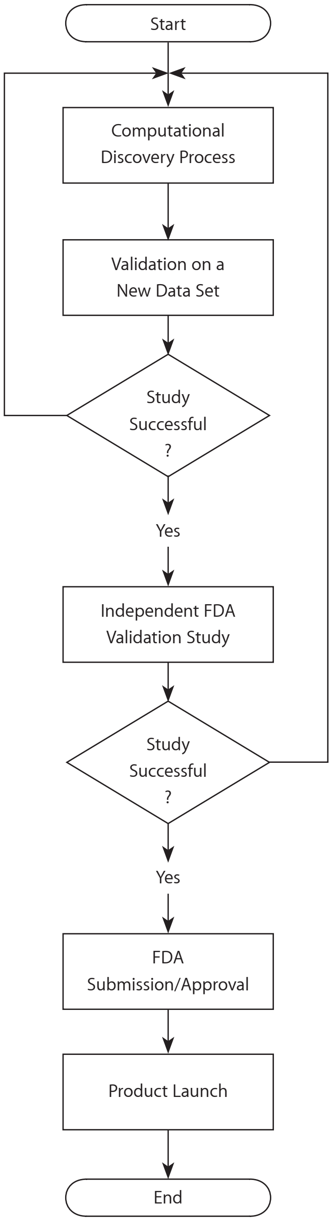
Before we begin our tour of the Exagen Discovery Engine, it will be helpful to view Exagen's methodological approach in the context of the alternative approaches that are available to researchers.

Deductive versus Inductive Research Methodologies

The deductive experimental approach to scientific research is responsible for most of the advances of modern science. In the field of genomics, however, current circumstances dictate alternative approaches. Genomics has provided the biomedical research community with a very large number of newly identified genes for which little or no functional information exists. And we are just in the early stages of truly understanding how genes control biological functions. Faced with an overwhelming number of imperfectly understood genes and a great deal of uncertainty about how genes, their regulation, and their products interact to bring about specific biomedical outcomes, the classical, deductive scientific method becomes insufficient.

By contrast, a computational methodology provides an effective way to discover relationships between biomedical conditions and genomic features for which little or no a priori knowledge exists. Moreover, by discovering strong correlations between specific genomic features and biomedical conditions, a pure computational approach can provide a starting point for scientists to elucidate the complex underlying biology of the disease.

Differences between Computational Approaches


Start
Computational Discovery Process
Validation on a New Data Set
Study Successful ?
Yes
Independent FDA Validation Study
Study Successful ?
Yes
FDA Submission/Approval
Product Launch
End

The figure at left shows a typical workflow for a company like Exagen that uses a primarily computational discovery process to develop and market commercial predictive or prognostic tests according to FDA guidelines.

What many such companies have in common is a process that identifies genomic markers that can accurately differentiate sample data from tissue samples representing different biomedical categories (for example, differentiating malignant from benign tissue).

A typical approach is to evaluate gene expression data—generated using microarrays—for each of the approximately 30,000 human genes, and to rank individual genes based on their predictive or prognostic utility for various outcomes. The result is a list of genes ranked by individual accuracy, something like this:

Table 1: Ranking of individual genes by accuracy

Rank	Gene	Accuracy
1	Q	0.756
2	M	0.727
3	V	0.699
4	A	0.678
5	C	0.654

Next, because no single gene in these results is sufficiently accurate for a medical test, the researchers develop a test by combining the top individual performers.

With this type of simplistic ranking approach, it might be necessary to include a dozen or more of the top-scoring individual genes to reach a level of accuracy that is acceptable for a commercial test. Moreover, having dozens of genes in the assay may make the commercial test more difficult and expensive to administer.

Using the ranking approach, it is possible to achieve reasonable levels of accuracy by looking at how subsets of the top performers in a data set work together. Table 2 shows what such a ranking might look like if the combinations evaluated were limited to the 100 highest-ranking genes.

Table 2: Top-ranking gene combinations discovered using a single-gene search methodology

Rank	Gene Combinations	Accuracy
1	The 63 highest-ranking genes, including genes M (Individual Rank 2) A (Individual Rank 4) Q (Individual Rank 63)	0.860
2	The 12 highest-ranking genes, including genes M (Individual Rank 2) V (Individual Rank 3) H (Individual Rank 12)	0.855
3	The 5 highest-ranking genes, including genes M (Individual Rank 2) A (Individual Rank 4) C (Individual Rank 5)	0.802
4	The 4 highest-ranking genes, including genes M (Individual Rank 2) V (Individual Rank 3) A (Individual Rank 4)	0.785

This approach generally assumes that each gene is an equal contributor and that single contributors are additive. Even using information from such a large number of the most accurate individual genes, however, there is a good chance that such an approach will fail to provide a test with an acceptable level of accuracy.

Exagen takes another type of computational approach to identify the combined contribution of genes simultaneously. The Exagen discovery process evaluates the accuracy of gene combinations directly, without regard to the absolute ranking of individual genes.

Table 3 contains an example of how this might look in practice. (Individual gene rankings from the previous example have been added to the Gene Set column for clarification.)

Table 3: Same search shown in Table 2 but performed using the Exagen approach (Example)

Rank	Gene Combinations	Accuracy
1	A (Individual Rank 4) C (Individual Rank 5) L (Individual Rank 238)	0.903
2	M (Individual Rank 2) A (Individual Rank 4) Q (Individual Rank 63)	0.860
3	M (Individual Rank 2) V (Individual Rank 3) H (Individual Rank 12)	0.855
4	M (Individual Rank 2) A (Individual Rank 4) C (Individual Rank 5)	0.802

Whereas the gene combinations obtained using the rank-ordered computational approach (Table 2) varied in size from 5 to 63 genes, each of the combinations obtained using the Exagen approach (Table 3) consists of exactly three genes. Notice that the second-, third-, and fourth-ranking gene combinations of the Exagen discovery process (Table 3) provide the same accuracy as the top three gene combinations obtained using the rank-ordered computational approach (Table 2), but with fewer markers, by eliminating redundant genes that do not provide any additional information.

The most important point to notice, however, is that the highest-ranking solution from the Exagen discovery process outperforms any of the rank-ordered gene combinations obtained using the alternative computational approach, even though it consists of the 4th, 5th, and 238th most accurate genes (considered individually).

This hypothetical example clearly illustrates the principle that although an individual gene may have poor predictive power on its own, it should not be discarded on that basis alone. In combination with other genes, it may provide the highest predictive accuracy available. By recognizing and giving primary importance to the significance of gene combinations rather than individual genes, the Exagen process optimizes overall performance while minimizing the number of biomarkers needed.

The Exagen Process: Discovering Biomarker Combinations

The inputs to the Exagen Discovery Engine can consist of gene expression data, proteomic data, gene copy number, and clinical information about the samples tested, as well as other available information. These data are derived from a set of patient samples that have known clinical outcomes. The data from a subset of these samples are randomly selected as the "training data." These training data serve as the input data to the Exagen Discovery Engine and are essential for testing the output of their computational processes.

An important difference between Exagen and other research companies doing computational biomarker discovery is Exagen's ability to simultaneously mine different types of data or multiple data sets. These capabilities involve mapping the different data sets to each other and applying appropriate statistical weights to each data type. When the availability of homogeneous data is limited, the ability to mine disparate data sets concurrently allows Exagen to combine data from two or more sets of samples, effectively increasing the size of the input data set and thereby improving the statistical significance of the computational results.

Exagen's computational scientists perform the mapping and weighting of data before the start of the computation process. At the end of any required mapping and weighting process, the input data set is ready to be fed to the discovery engine along with the initial processing parameters.

One key step in the process is the selection of a classifier or classification algorithm that is appropriate to the input data and the desired result. Additional parameters are also selected to further guide the search. For example, in most cases, there are too many possible combinations to do an exhaustive search, so there are parameters that specify early on whether an exhaustive search is possible or a heuristic search is necessary. A heuristic search is one that is not exhaustive but instead narrows down the tested combinations by, for example, looking broadly across possible combinations and concentrating the search process on genomic regions where the results are most promising.

After setting the operating parameters, the discovery scientist runs the search process. The search process consists of an iterative test on successive combinations of genomic features and a scoring mechanism, called the *objective function,* that is assigned to each combination based on how accurately that combination was able to classify the training data. A typical search process consists of 10,000 to 50,000 of these iterations.

While the process is running, the system operator can monitor its progress and make adjustments to the operating parameters to provide external "steering" to the discovery engine. For example, if the operating parameters are originally set to search for combinations of 10 features but the input data supports significant results with combinations of only three features, the operator can quickly determine that the results are unsatisfactory and can adjust the parameters accordingly.

The search algorithm is a coarse-grained problem; that is, one that can easily and efficiently be divided into computationally independent subtasks and processed in parallel on multiple computers. This is where Exagen's dual processor Xserve G5 computational cluster pays big dividends. The combinations to be tested are divided into batches, each of which is assigned to a different node on the cluster. After all of the data has been processed and reassembled, the tested combinations are ranked according to their objective functions and the search phase of the computational discovery process is complete.

The Exagen Process: Significance Testing

At this point, the discovery engine has identified thousands of marker combinations, several of which will subsequently be validated in an independent study required for FDA clearance. Before proceeding to that study, however, Exagen adds an in silico prevalidation phase, called *significance testing,* to the computational discovery process.

All data sets contain accidental relationships that can introduce errors into the results of the search algorithm. The smaller the sample set (in this case, biological samples collected from patients) in relation to the number of features (in this case, genes or genomic areas), the higher the probability of error. One problem that is common to all genomic research facilities is that the number of features is often large (on the order of 30,000 genes) and the number of samples is almost always too small by comparison. As a consequence, there is a built-in risk that the apparent accuracy of some marker combinations results from "noise" rather than meaningful information.

The validation studies required for FDA clearance detect such anomalies by collecting new data samples and validating the effectiveness of the proposed solutions against that new data. However, the FDA-mandated validation trial is so consequential—in time, money, and the diverted attention of key personnel—that Exagen goes to extraordinary lengths to weed out weaker results computationally before sending proposed solutions into independent verification.

For the purpose of illustration, imagine a situation in which the company is searching for biomarker combinations that can distinguish samples taken from prostate cancer tumors from samples taken from normal tissue. The significance testing routine begins with data that is correctly labeled as corresponding to either malignant or nonmalignant tissue, as shown in Table 4.

Table 4: Original data sets, correctly labeled

Sample Number	Biomedical Condition	Gene 1	Gene 2	Gene 3	Gene 4	…	Gene 30,000
Sample01	Disease	data11	data 12	data 13	data 14	…	data 130000
Sample02	Disease	data 21	data 22	data 23	data 24	…	data 230000
Sample 03	Normal	data 31	data 32	data 33	data 34	…	data 330000
Sample 04	Normal	data 41	data 42	data 43	data 44	…	data 430000

The first step is to randomize those labels, as shown in Table 5 (column highlighted in yellow).

Table 5: Data sets with "Biomedical Condition" column randomized

Sample Number	Biomedical Condition	Gene 1	Gene 2	Gene 3	Gene 4	…	Gene 30,000
Sample 01	Disease	data 11	data 12	data 13	data 14	…	data 130000
Sample 02	Normal	data 21	data 22	data 23	data 24	…	data 230000
Sample 03	Normal	data 31	data 32	data 33	data 34	…	data 330000
Sample 04	Disease	data 41	data 42	data 43	data 44	…	data 430000

The next step is to run the discovery process using the intentionally randomized data, once again attempting to identify the gene combinations that perform best in terms of accuracy. This process of re-randomizing the labels and re-running the discovery process is repeated as many times as feasible, collecting the results each time. Typically, 50 iterations provide a reasonable balance between cost and confidence, yielding a p-value on the order of 0.02.

Each discovery process produces a set of top-performing gene combinations that can then be compared, as shown in Table 6.

Table 6: Example of significance testing results

Results	Member 1	Member 2	Member 3	Member 4	Accuracy
Best Combination from Original Discovery Process	Gene	Gene	Gene	Gene	0.91
2nd-Best Combination from Original Discovery Process	Gene	Gene	Gene	Gene	0.89
▲ Results probably significant ▲			▼ Results probably not significant ▼		
3rd-Best Combination from Original Discovery Process	Gene	Gene	Gene	Gene	0.77
Best Combination from Any Randomization	Gene	Gene	Gene	Gene	0.75
4th-Best Combination from Original Discovery Process	Gene	Gene	Gene	Gene	0.73
2nd-Best Combination from Any Randomization	Gene	Gene	Gene	Gene	0.72
3rd-Best Combination from Any Randomization	Gene	Gene	Gene	Gene	0.63

The accuracy of candidate combinations resulting from the original discovery process can now be evaluated in relation to a benchmark established by the best results using randomized data. Only those results that are markedly more accurate than the best results from randomized data are considered to be significant. In Table 6, for example, the best and second-best results from the original discovery process (accuracy of 0.91 and 0.89, respectively) are clearly better than the best results using randomized data (accuracy of 0.75). The third-best combination from the original discovery process, with an accuracy of 0.77, is probably not significant because it is only marginally more accurate than the results using randomized data.

The computational cost to run the Exagen Discovery Engine using a set of randomized data is identical to that of the original discovery process. A typical significance testing regimen comprises 30 to 50 such runs and is, therefore, 30 to 50 times as costly as the original discovery process.

This completes the computational discovery process and our tour of the Exagen Discovery Engine. The marker combinations that survived the significance tests can now be validated in a clinical study. From this point forward, the Exagen process is exactly the same as that for any company pursuing FDA approval for its predictive or prognostic tests to become in vitro diagnostic (IVD) products.

Conclusion

The idea that a complex disease such as cancer occurs solely because one gene is overexpressed is not supported by the evidence, even though it is theoretically possible. Instead, the evidence is mounting that complex diseases are the result of events that occur in multiple pathways. By developing computational tools to detect the involvement of multiple interacting genes in specific medical outcomes, Exagen can provide solutions that go beyond the capacity of any single-gene solution or ranking of genes one by one.

The multiple-gene solutions discovered using the Exagen process provide immediate benefits by allowing caregivers to deliver better diagnostics and treatment decisions for their patients and by helping pharmaceutical companies to develop new and better treatments.

In addition, by identifying combinations of genes that work together to bring about specific biological outcomes, the research being done at Exagen helps to identify promising genes for future research and, potentially, new drug targets to improve the quality of the therapeutic tools available to physicians for fighting disease.

The real strength of the Exagen discovery process, and of the primarily computational approach it represents, is precisely that it does not require any foreknowledge of specific gene functions or gene interactions in order to produce meaningful results. On the contrary, it is a powerful tool to help scientists to explore previously uncharted genomic territory—a tool that would have been inconceivable without the enormous computational power of high-performance computing platforms such as the dual processor Power Mac G5 and Xserve G5.

Apple computers have for many years played a prominent role in academic research computing. Now the reliability and ease of use of the UNIX-based Mac OS X Server, together with the power and affordability of the Xserve G5 and Xserve RAID hardware, are making Apple computers the computing platform of choice for leading-edge scientific computing companies such as Exagen.

Visit the Apple Science website at www.apple.com/science to learn more about Apple's high-performance computing solutions.