

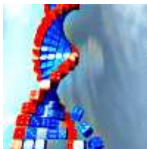
Project Room: Challenge Details

PROJECT ROOM MY SOLUTIONS TEST SOLUTION MESSAGES(7) FORM A TEAM

Predictive Data Analysis

Challenge Reward: **\$100,000 USD** Challenge Type: **RTP**

INNOCENTIVE 9231572



The Seeker is looking for Solvers to analyze a large set of data. The Seeker believes that creative Solvers that bring unique or novel computational means may provide the most exciting and high performing solutions. *You can try our new **Team Project functionality** on this Challenge.*

 DEADLINE: Jul 26, 2010

 1850 Project Rooms

 Posted: Feb 27, 2010

This is your secure and confidential Project Room for the Challenge. From here, you can receive the Challenge details, submit your solution proposal, ask questions, and receive answers confidentially from the InnoCentive team.

Detailed Description & Requirements

Background & Impact Statement

NOTICE: Ineligibility to Participate:

Employees or contractors of the following organizations are ineligible to participate in this Challenge and may not access the Training or Testing Datasets nor may they use the Prodigy:

- Asgrow
- BASF
- Bayer
- Dekalb
- Dow
- Dupont
- Monsanto
- Pioneer
- Syngenta

If you are an employee or contractor to any of the companies listed above, **EXIT THE CHALLENGE AT THIS POINT.**

The analysis of digital data has long been an area of focus for the computational sciences. However, due to the fractured and variegated nature of the applications served by the computational sciences, it is highly likely that sophisticated methods exist within one computational domain but have yet to be applied to solve problems within other domains. This Challenge seeks to make connections between various computational domains by focusing a diverse audience of Solvers onto the interpretation of a tremendous amount of digital data.

This Challenge asks Solvers to develop and apply both known and novel methods of interpreting a large digital dataset. Solvers from any sort of computational background including electrical engineering, computer science, mathematics, physics and bioinformatics are encouraged to participate.

The Seeker strongly believes that other disciplines may have valuable ideas to contribute to problems in computational biology. Although the desired solution can be expressed as a pure data-mining problem (see equation 1 below), we believe the some understanding of the biology may help Solvers refine their solutions. Therefore if you wish to have more background information on the biology, please refer to the section entitled **Basic Biological Background Information** for a slightly simplified yet sufficient introduction to the essential elements of the biology contained within this problem.

The Challenge is focused on the interpretation of digital representation of the DNA found in many breeds of an organism. The three inputs into the problem (the X variables) are (A) a reference DNA sequence of the organism, (B) single-letter variations (deviations) from the reference sequence for each breed and (C) gene (mRNA) expression (activation) levels for each gene for each breed. Ultimately the

The Submission Process

To submit your solution proposal to InnoCentive, please click the "Submit Your Solution" button.

[Submit Your Solution](#)

Answering Your Questions

If you have questions regarding this Challenge click the "**Messages**" button.

[Messages](#)

Solver Agreements

These are the Solver Agreements you have signed for this Challenge.

 [Solver Terms of Use](#)

 [Challenge-Specific Agreement](#)

Additional Information

[Frequently asked questions](#)

Seeker would like the Solvers to use these data to model an observed Y variable which represents an important trait to the organism.

Specific Description of the problem

Dataset A (The Reference Genome) is a set of strings of letters from a four-letter alphabet (A, T, G and C). This is the reference genetic sequence for the organism and each string is the sequence for a particular *chromosome from which thousands of genes exist as discreet units of sequence along the chromosome*. However each breed of the organism has a genetic sequence that differs from the reference sequence at several positions. Those individual variations for each breed are recorded in Dataset B (Variant). Our hypothesis is that the difference in trait Y between the breeds is partly caused by these individual variations in the sequence. Thus the individual variations among breeds at a particular position on the string A can be considered to be potentially predictive of Y.

Genes act as template for producing mRNA which in turn is used as a template for producing protein. In Dataset C (Expression) we provide the abundance of mRNA from many genes for each breed. Also in Dataset C the discreet location of each gene relative to the reference genome is reported (Maskedgene_positions) to facilitate integration of the Variant and Expression datasets. The relative abundance of mRNA for a particular gene can also be considered to be potentially predictive of Y.

Using the reference sequence (Dataset A) and the individual variations recorded in Dataset B, it is possible to reconstruct the genetic sequence for each breed. Let us denote G_{bi} as the sequence at the i^{th} position and E_{bj} , the expression of the j^{th} gene for the breed b . G_{bi} and E_{bj} are hypothesized to be predictive of Y_b , the value of the trait Y. For some selected sequence variations $i \in I$ and expressions of genes $j \in J$ the trait Y can be modeled as

$$Y_b = f(G_{bi}, E_{bj} \mid i \in I, j \in J)$$

The model function might also depend on the context of G_{bi} , i.e. the sequence flanking the i^{th} variation.

The organism trait Y was tested in two distinct environments and at two different times, giving 4 independent measurements of the trait. The Seeker has measured the trait Y and ranked the organisms in accordance with its mean trait Y. It can be assumed that all samples in the dataset were grown in similar growing conditions.

Problem difficulty

The difficulty of this problem is that, as with most computation biological problems, it is highly under-constrained in that the number of variables in X far exceeds the number of observations in Y. There are >1,000,000 potentially predictive variations G_{bi} in the genetic sequence (Dataset B) and >20,000 potentially predictive gene-mRNA abundances E_{bj} in Dataset C. In addition, the two data-types are of different kinds: for each breed G_{bi} takes discrete values from a four letter alphabet while E_{bj} is a real-valued number. So methods that can integrate heterogeneous data (e.g. Bayesian classifier, Voting methods etc.) may outperform methods that operate only on Dataset B or Dataset C. Solvers are not required to use both datasets in their answers nor are they required to use all the data in each dataset.

The Seeker strongly believes that some progress may be made towards the ultimate goal of predicting Y given the Datasets A, B and C.

Objective

The objective of this Challenge for the Solvers is to apply novel methodologies to use the X data to rank the measured Y, provided in Dataset D (Phenotype), from highest to lowest predicted output. For example if there were 10 samples in a subset of the data, then the ranked list: 9, 7, 10, 2, 6, 4, 1, 3, 5, 8 indicates that sample 9 has the highest predicted output and sample 8 has the lowest predicted output.

Any method may be used as long as it is reproducible and reasonably computable by the Seeker on commodity hardware (standard PCs or servers). Algorithms that require substantial investments in additional computing power are potentially interesting but obviously less attractive to the Seeker.

The Seeker has provided Solvers with a complete training dataset which includes molecular profiling information (X) and trait data (Y) on 100 different breeds of the organism. This includes Dataset A, Dataset B, Dataset C and Dataset D. The datasets may be downloaded below in the **Dataset Download** section.

The objective is for Solvers to produce the most accurate model for predicting the trait Y based solely on the data provided in the Training Set of data and any publically available external data resources. Eventually the top Solvers models will be evaluated on additional equivalent datasets that have not been shown previously to the Solvers. This implies that over fitting of any model or method should be avoided at all costs. Employing standard means of cross-validation are strongly recommended. To aid Solvers in the development of their solution, InnoCentive has created a solution testing featured known as the Prodigy and described below.

Solvers which submit solutions which have an R^2 exceeding ~~0.44~~ 0.30 (revised, please see May 18th Challenge Addendum below) will be eligible for the full award after it has been demonstrated on an independent test set provided by the Seeker during final evaluation.

Solution Testing Tool – The Prodigy

(new InnoCentive website feature available after March 8th)

The solution testing feature known as the Prodigy allows Solvers to try their algorithm on an independent *Testing* dataset and displays a table of the top 10 scores achieved by participating Solvers since the beginning of the Challenge. InnoCentive believes that this feature will stimulate better competition among the Solvers and make for a more engaging and successful Challenge.

Use of the Prodigy does NOT constitute a valid and complete submission. Solvers *must* complete a typical solution submission using the 'My Solutions' tab above. Only those solutions that are made through the My Solutions tab will be evaluated by the Seeker. Solvers

Project Room - Challenge Details

are encouraged to include their score as measured by the Prodigy within the solution documentation they submit via the My Solutions webpage. Please see the **Deliverables** section below for details on what the submissions should contain.

Because of its novelty, below we have outlined the details and terms of use for the Prodigy.

Input Data:

The Seeker has provided an independent test dataset of 149 breeds of the organism. Like the Training Set, the Test Set includes Datasets B and C. The same Dataset A from the Training Set is used for the definition of Dataset B in the Test Set.

The Test Set, unlike the Training Set, does not include the trait Y values. Solvers are asked to use their proposed methods to predict the relative trait values Y of the 149 breeds of the organism in the test set.

The input into the Prodigy is a ranking of the 149 Testing Set breeds from highest to lowest predicted yield. In the Y values file (phenotype), the test set samples each have an assigned Index of 101 – 249. Consider, for example, the following input:

210, 130, 234, 212, 111, 202, 103, 117, 211, 134, 162, 119, 218, 238, 141, 209, 200, 197, 183, 101, 247, 186, 114, 104, 207, 169, 142, 176, 228, 193, 227, 167, 184, 225, 226, 185, 198, 147, 166, 125, 108, 143, 240, 192, 152, 118, 171, 154, 126, 155, 230, 248, 127, 187, 178, 110, 132, 116, 112, 172, 113, 168, 148, 233, 146, 107, 217, 106, 199, 249, 161, 205, 129, 173, 157, 124, 231, 220, 213, 203, 164, 237, 190, 115, 170, 206, 214, 181, 196, 105, 195, 194, 243, 139, 232, 131, 175, 177, 180, 242, 122, 235, 208, 144, 182, 236, 159, 121, 219, 246, 244, 138, 165, 128, 201, 145, 245, 160, 102, 189, 150, 135, 140, 229, 120, 123, 239, 221, 163, 216, 204, 174, 137, 222, 223, 241, 151, 153, 156, 109, 224, 191, 188, 133, 136, 179, 149, 215, 158

This indicates that the breed with index number 210 has the highest output (predicted Y), the breed with the index 130 has the second highest output and solution breed index 158 has the lowest output. Note that the interpretation of this "ranking" input differs from the output of programming languages such as R which represents rankings differently. The above format is more analogous to the output from an 'order' function in some programming languages.

The input to the prodigy tool must contain only the unique set of numbers between 101 and 249 in a comma separated format that is equivalent to the above format. No other inputs will be accepted.

Testing Methodology:

Upon entering the input data into the Prodigy, the website will instantly compute a [Spearman's Rank Correlation](#) between the Solver's input and the known (measured) trait data that the Seeker has provided to InnoCentive. The score is displayed in the yellow box as the Spearman's correlation coefficient squared (r^2). The website maintains at least 8 digits of accuracy but displays only 4.

Due to the imprecision of scoring solutions composed of only 149 observations, r^2 values that are less than 0.14 are not statistically significant. Therefore, only scores equal to or exceeding 0.14 will be shown in the leaders table.

Display of Username:

Should a Solver submit an ordering that has a correlation exceeding 0.14 and within the top 10 highest scores achieved at that time, the Solver's name will be displayed in the leaders table. The leaders table will be continuously updated, so Solvers are encouraged to check in frequently to see how Solvers are making progress toward the solution. InnoCentive cannot remove the leaders table entries of Solvers. By using the tool you are providing consent to InnoCentive to display your username and score.

Daily Use Limit:

In order to prevent gaming of the system, InnoCentive has imposed a limit of 5 submissions per Solver per day to the Prodigy. Each new day starts at 12:00AM Eastern US time.

Abuse of the Tool:

InnoCentive will not accept the misuse of the Prodigy. InnoCentive reserves the right to revoke the privilege of Solvers that misuse the tool for anything other than its intended purpose. Intentionally malicious inputs, including repeated submission of solutions that appear to reverse engineer the solution, constitute abuses of the tool. InnoCentive may choose to suspend the account or accounts of any Solver that is suspected of misuse.

Final Submission Evaluation:

Upon the conclusion of the Challenge, the Seeker will evaluate only the submissions which have been made on the My Solutions page. Use of the Prodigy does NOT constitute a submission and will not be evaluated.

The Seeker reserves the right to award any of the solutions submitted. The Seeker's formal evaluation process will consider the quality of documentation, algorithmic efficiency and opportunity for future expansion among other solution attributes. The final awarded winners may or may not be represented on the leaders table.

Accessing the Prodigy

The Prodigy tool will be made available to Solvers after March 8th. At that point Prodigy can be accessed by using the Test Solution tab at the top of this page and circled in the image below.



Dataset Download

Project Room - Challenge Details

Both Training and Testing Sets have been broken up into chunks of data to facilitate downloading.

There are 7 data files for download of which 2 are subdivided into 11 files:

- [Variant_Training.zip](#) (subdivided into 11 files representing CH0-CH10)
- [Variant_Test.zip](#) (subdivided into 11 files representing CH0-CH10)
- [Expression_Training.zip](#)
- [Expression_Test.zip](#) [Updated Expression Test Data](#) *Please see April 23rd Challenge Update below for more details.*
- [Phenotype_Training.csv](#)
- [Phenotype_Test.csv](#)
- [Maskedgene_positions.zip](#)

"WARNING: Should you choose to open the variant data files in MS Excel, when using the text import wizard be sure to specify in step 3 that each column be imported in "text" format. If this is not done, field values will be corrupted as evidenced by the appearance of "000" values."

The Reference Genome data set is named version 4a.53 and can be obtained online at <http://ftp.maizesequence.org/release-4a.53/>. All references to the single letter genomic variations in both the Training and Test datasets are built off of this reference sequence. Variants located in non-coding regions of the genome were not included in this dataset but not all coding regions are captured in the MaskedGene_Positions dataset so some gene variants may be outside the defined gene positions.

Data Type Descriptions

To ensure proper interpretation of the datasets, they are briefly described here.

DatasetA: Reference Genome

- This dataset represents the reference genetic sequence for all chromosomes each numbered between 1 and 200-300 million based on the number of DNA base pairs in each chromosome. In linear order.
- The data are publically available via the web with various levels of annotation from <http://ftp.maizesequence.org/release-4a.53/>.
- The name of the reference genome is 4a.53. A new version may be available by the end of April 2010 so please ensure that 4a.53 is accessed for any use with the Challenge

DatasetB: Variant

- A matrix of differences between the sample and the Reference Genome at specific nucleotide base pair position. The exact difference is not provided.
- Column Headers:
 - Difference: formatted to show the numeric location and the chromosome of the difference. 12345CH1 = Nucleotide position 12345 on Chromosome 1.
 - Type:
 - S = SNP (Single Nucleotide Polymorphism) difference and specified location.
 - InDel = An insertion or deletion difference starting at the specified location. Sizes of the insertion or deletion are not provided.
 - Items_with_difference: contains only sample with high quality data supporting a difference with the reference exists at the specified location.
 - Items_with_no_difference: contains only samples with high quality data supporting no difference with the reference exists at the specified location.
 - Items_with_ambiguity_favoring_difference: contains only sample with variable quality data suggesting a difference with the reference might exist at the specified location.
 - Items_with_ambiguity_favoring_no_difference: contains only samples with variable quality data suggesting no difference with the reference might exist at the specified location.
 - Items_with_no_information_at_this_position: contains all samples that have no data at the specified location.
- In all cases, the location of the difference on each chromosome is in linear order so the location of each difference relative to the other differences is known. For example 12345CH1 is exactly 50 nucleotide base pairs away from 12395CH1.
- Chromosome 0 (CH0) is the designation for known genome sequence that is currently not part of the linear order of the 10 other chromosomes. CH0 is a collection of appended sequences and is therefore not in a linear order of sequence.
- The variant files for the Training and Testing data sets are subdivided by Chromosome into 11 files each in order to keep the file sizes small for easier download.

Dataset C: Gene Expression

- An absolute expression level for known genes for each sample.
- The expression data was acquired from a single time point (2 weeks after germination) from a single tissue type (Leaf). The expression data was not derived at the same time or under the same conditions as the phenotypic data output. Each column of data is annotated with 'X' and the corresponding sample identifier in the phenotype files.
- Not all genes will have expression values because:
 - the gene is not expressed in that tissue at that time
 - the expression level was too low to be detected by the method

Project Room - Challenge Details

- Solvers are encouraged to determine their own threshold for distinguishing background noise from accurate expression signals.
- A reference data file [Maskedgene_positions](#) is included to help solvers define gene space along the Genome Reference (available online) and resolve gene specific differences as defined in the Variant data file.

Dataset D: Phenotype

- Column Headers:
 - Index: The Challenge unique identifier for each sample
 - Identifier: The Sample unique identifier that links to the expression and variant data sets
 - DataSet: The designation for samples belonging to either the Training (Train) or Test datasets. The Test dataset does not have data for phenotypic output since the prediction of the mean output ranking is the objective of this Challenge.
 - Feature Group: Three categories representing a high level biological based difference between the samples.
 - Output_loc1_Year1: phenotypic output from Location 1 in Year 1 for each sample
 - Output_loc2_Year1: phenotypic output from Location 2 in Year 1 for each sample
 - Output_loc1_Year2: phenotypic output from Location 1 in Year 2 for each sample
 - Output_loc2_Year2: phenotypic output from Location 2 in Year 2 for each sample
 - Output_Mean: Mean of 4 phenotypic output data points
 - Output_STDEV: Standard deviation of mean from 4 phenotypic output data points

Rank Based on Output Mean: Rank of each sample from high to low based on Output_Mean

Deliverables:

Solvers are asked to submit proposals that include the following information:

- A thorough and detailed description of the Solver's algorithm and/or approach.
- Well documented or commented source code implemented in R, Java, Python, Matlab, C, C++ or SAS
- A list of the Test Dataset's 149 breeds ordered from highest to lowest predicted trait Y. This should be included directly into the submission document. The associated score as determined by the Prodigy would also be useful for the Seeker's evaluation.
- References to appropriate concepts, data or information used by the Solver.

Results from the Prodigy will not be the only means of evaluating the submissions.

Technical Requirements:

In order to be eligible for an award a solution will need to exhibit a performance gain that is beyond chance ($r^2 > 0.14$). During the full evaluation process, Solvers will be asked to run their algorithms on additional datasets are not provided at this time.

Additional considerations will also be made concerning the solution's ease of computability, quality of documentation and potential for future improvement.

The winning solution may or may not be represented on the Prodigy's leader's table.

Basic Biological Background Information:

Cells make up all living organisms. The behavior of the population of cells in an organism is what determines the behavior and appearance of the organism as a whole. All cells contain a long sequence of a four letter (A, T, C and G) code known as DNA. The DNA sequence contains many discrete (although sometimes overlapping) regions called genes. The letters within each gene region encode RNA (a complementary four-letter code) and the RNA is then used to encode proteins. Proteins in turn serve as the workhorses of the cell and carryout just about all vital cellular functions. Ultimately vast the majority of a cell's minute-to-minute activities are performed by use of its proteins. Biology's control of the proteins is largely carried out by turning genes on or off (altering each gene's *expression*) depending on the cells' need for the proteins they encode.

All cells within an organism have the same DNA code. However cells in different organisms have different genetic codes. It is a result of the genetic sequence differences as well as the impact of environmental forces on those genetic sequence differences that causes diversity among organisms of the same species. By and large it is believed that a substantial proportion of the differences between two organisms can be explained by differences in the genetic code or some downstream manifestation of that code (e.g. RNA or proteins).

The cells within organisms display differences in baseline behaviors, response to stimuli, and physical appearance as a result of several factors. The factors explored within this Challenge are outlined below:

- There may be variations in the DNA sequence of two different organisms (Dataset B). This leads to different gene sequences and slight differences in the resulting protein created by the two organisms.
- For two different organisms of the same species genes are activated in slightly differently patterns (Dataset C). This difference can be observed by differences in the RNA levels of the proteins.
- Other differences certainly exist but for the purpose of this Challenge, those two variables are the most important ones to consider.

April 23rd Challenge Update

It has come to our attention that there was a slight issue on the normalization of the testing data for gene expression. Therefore, the Seeker had released an updated data file to replace the existing one. We encourage you to download this file and include it in your analysis.

The updated file is provided [here](#).

Additionally, the Seeker understands the complexity of the genotypic data and difficulty inherent to training predictive algorithms on data sets with unbalanced or missing data. In order to avoid pre-supposing a solution on the Solver community the data were posted in its most basic form. Had this been done in-house, the Seeker may have normally perform an imputation process to improve the balance of the data set prior to further analyses. References and reviews for some common imputation methods are included below if the Solver's are interested in exploring this approach to the challenge.

The Seeker recognizes that this may not be a necessary step to solve the challenge and makes no recommendation towards the relative merit of different imputation methods. Furthermore, other imputation methods are likely also available online and may provide benefits that these methods do not.

[Missing value imputation for epistatic MAPs; BMC Bioinformatics. 2010 Apr 20;11\(1\):197. Ryan C, Greene D, Cagney G, Cunningham P.](#)

[Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies; BMC Genet. 2009 Jun 16;10:27. Hao K, Chudin E, McElwee J, Schadt EE.](#)

[Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows; Bioinformatics. 2007 Jul 1;23\(13\):i401-7. Roberts A, McMillan L, Wang W, Parker J, Rusyn I, Threadgill D.](#)

[A flexible and accurate genotype imputation method for the next generation of genome-wide association studies; PLoS Genet. 2009 Jun;5\(6\):e1000529. Howie BN, Donnelly P, Marchini J](#)

[A comprehensive evaluation of SNP genotype imputation; Hum Genet. 2009 Mar;125\(2\):163-71. Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A.](#)

May 18th 2010 Challenge Addendum:

The Seeker has decided to lower the awarding threshold from an $r^2 \geq 0.41$ to $r^2 \geq 0.30$. Therefore, Solvers who submit solutions which have an R^2 exceeding 0.30 will be eligible for the full award after it has been demonstrated on an independent test set provided by the Seeker during final evaluation. Similarly, the leader's table will now show submission which have a $r^2 \geq 0.12$.

Test Your Solution:

For this Challenge, you will be able to test your algorithm before submitting your final solution. Please use the link below to access the test area.

[Test your Solution](#)

Project Criteria

Submissions will be judged by means of the documentation submitted through the My Submissions website interface for this Challenge. **Submissions made solely to the Prodigy will not be evaluated by the Seeker and do not constitute an eligible submission.**

A winning solution will have the following characteristics:

- Well documented algorithm with clear explanation of its function and computational underpinnings.
- Well documented and demonstrative software code that illustrates the Solver's proposed algorithm.
- A high degree of performance found by using the Prodigy of equivalent means. The Seeker believes that a solution eligible for a full award will have an R^2 exceeding 0.44-0.30 (revised, please see May 18th Challenge Addendum above).

During the evaluation process, the Seeker reserves the right to ask some Solvers to demonstrate their algorithms on additional datasets that are not distributed with this Challenge.

Team-based Proposals

We value the diverse nature of the Solvers in our Network, and are now encouraging you to strengthen your Proposals by recruiting team members to work on this Challenge. Past experience shows that collaborating with multi-disciplinary colleagues and submitting Proposals as a team can truly yield great results.

To support team collaboration, we have added new functionality called a "Team Project Room". A Team Project Room is a secure online workspace that allows a group of Solvers to securely collaborate and solve an InnoCentive Challenge. Team Project Room functionality will only be available for selected Challenges. By encouraging Solvers to work together, we believe that the quality and quantity of

Project Room - Challenge Details

solutions to more complex or multidisciplinary Challenges will be improved. If you want to read more about Team Project Rooms, [click here](#).

Here's how to do it.

1. Find team members and have them register for a Solver account on InnoCentive.com
2. Once you have your team, click on the Form A Team tab found in the project Room for this Challenge and fill in the required fields.

That's it!

All inquiries will be responded to within 1 business day. We look forward to seeing your teams collaborate in our new work environment, and would greatly appreciate any feedback.

Project Room | [My Solutions](#) | [Messages\(7\)](#) |

[Home](#) | [Products](#) | [Seekers](#) | [Solvers](#) | [Challenge Center](#) | [RSS Home](#) | [My InnoCentive](#) | [Help](#) | [About Us](#) | [Terms of Use](#) | [Privacy Policy](#)

Copyright©2010 InnoCentive, Inc. All rights reserved.