

Bioinformatics eLearning

Jason H. Moore, Ph.D.

Third Century Professor

Professor of Genetics

Professor of Community and Family Medicine

Director, Institute for Quantitative Biomedical Sciences

Associate Director, Norris-Cotton Cancer Center

The Geisel School of Medicine

Dartmouth College

www.epistasis.org

jason.h.moore@dartmouth.edu

Bioinformatics Challenge Areas

- Collaboration – who has the time?
- Core facilities – service vs. research
- Data – too much!
- **Education – multidisciplinary training**
- Funding – strategies
- Hardware – grid computing
- Research – chasing technology
- Resource sharing - networks
- Software – biology vs. computer science
- Visualization – the future

WE SUGGEST YOU PROCEED IN THIS SEQUENCE ↓

Overview

Scientific Question

Biological Data

What Is Data Mining?

What Is Machine Learning?

What Is A Model?

How Good Is A Model?

Decision Tree Models

Decision Tree Analysis

Behind The Scenes

Beginning of
Module**Overview**Scientific
Question →

Welcome to the NH-INBRE Bioinformatics Education Module on **Data Mining**. The goal of this module is to introduce the basic concepts of data mining and machine learning. These bioinformatics methods are central to many biological and biomedical problems. We will use the disease tuberculosis (TB) as a motivating example.

We recommend you begin with the Scientific Question section and then proceed once you have a good understanding of the material. Some sections have a link to Wikipedia where you can do some additional reading and find links to other web pages.

Once you understand the basic concepts you can then try the example data mining problem. The TB data are from real human subjects and the machine learning method you will try is widely used in bioinformatics.

This module currently works with recent versions of the Firefox, Safari and Chrome browsers. The critical section **Decision Tree Analysis** does not yet work with Internet Explorer.

← Overview

Scientific Question

Biological Data →

Can the development of tuberculosis (TB) be predicted by cytokine levels?



TB is an infectious disease that is caused by different strains of *Mycobacterium tuberculosis*, a type of bacteria. The TB bacteria infect the lungs resulting in cough, chest pain, fever, weight loss and other symptoms such as fatigue. In 2007, more than one million people from around the world died from TB. For more information about TB visit [Wikipedia Tuberculosis page](#).

The goal of this study is to determine whether the development of TB can be predicted by cytokines. Cytokines are proteins that are used by the body to send signals to cells to change their function. For example, the Interleukin 4 (IL-4) cytokine helps activate cells that promote the healing of wounds. Many cytokines target cells in the immune system and are thus important for understanding how the body responds to infections such as TB. For more information about cytokines visit [Wikipedia Cytokine page](#).



This is a research study of approximately 2500 African subjects infected with human immunodeficiency virus (HIV). A total of 1181 of these subjects received an experimental vaccine aimed at preventing a secondary infection with TB. Subjects were followed over time to determine whether they develop a probable or definite TB infection. Subjects are labeled 1 if they develop TB or 0 if they do not. A total of 597 subjects developed TB. Of these, 210 were vaccinated. Table 1 below summarizes these numbers. The presence or absence of TB is the biological endpoint we are interested in predicting.

	TB	No TB	Total
Vaccinated	210	971	1181
Not Vaccinated	387	933	1320
Total	597	1904	2501

A total of 27 different cytokines (e.g. il1b, il4, ccl3, vegf, etc.) were measured in all of the subjects. In addition, the data set includes variables indicating whether the subjects were vaccinated (0=not vaccinated, 1=vaccinated), their age (in years) and their gender (0=female, 1=male). Table 2 is a made up example of what the data set looks like (only a few cytokines are shown). It is important to note that the human subjects used in this study have had their personal identifiers removed and are thus anonymous.

Subject	vaccine	il1b	il1ra	il2	il4	age	gender	TB
1	1	3.21	99.23	6.56	0.51	20.17	0	0
2	0	4.02	230.12	0.70	1.32	44.73	0	1
..
2501	1	1.22	49.94	10.77	0.91	34.01	1	1

← Biological
Data

What Is Data Mining?

What Is Machine
Learning? →

The goal of **data mining** is to identify patterns across multiple variables (also called **attributes** or **features**) in large data sets that help answer a scientific question. We are assuming here that the development of TB is a complex process that depends on multiple cytokines and other variables such as age, gender and whether the subjects were vaccinated for TB. For more information about data mining visit [Wikipedia Data Mining page](#).



← What Is Data Mining?

What Is Machine Learning?

What Is A Model? →

Machine learning methods are often used to mine complex patterns from large data sets. Machine learning uses computer algorithms to **learn** which combinations of variables are predictive of an **endpoint** (also called **class**) such as TB. Machine learning often has two objectives. The first goal is to select a subset of the variables that will be analyzed. This is referred to as **variable selection**. The second goal is to **fit a model** to the data. For more information about machine learning visit [Wikipedia Machine Learning page](#).



← What Is
Machine
Learning?

What Is A Model?

How Good Is A
Model? →

A model describes the relationship between the values of the variables and the endpoint. For example, a model of TB could be


```
IF il2 ≥ 5.485 AND il8 ≥ 171520 AND il8 < 227752 THEN TB=1 ELSE TB=0
```

This model says that you would classify a subject as having TB if their il2 levels were greater or equal to 5.485 and il8 levels at least as high as 171520, but less than 227752. Otherwise, you would classify them as not having TB. A computer algorithm would fit a model to the data by trying different combinations of variables along with different mathematical functions (e.g. <, >, =) until it finds a particular model that does a good job of predicting who develops TB. For more information about models visit [Wikipedia Mathematical Model page](#).

← What Is A Model?

How Good Is A Model?

Decision Tree Models →

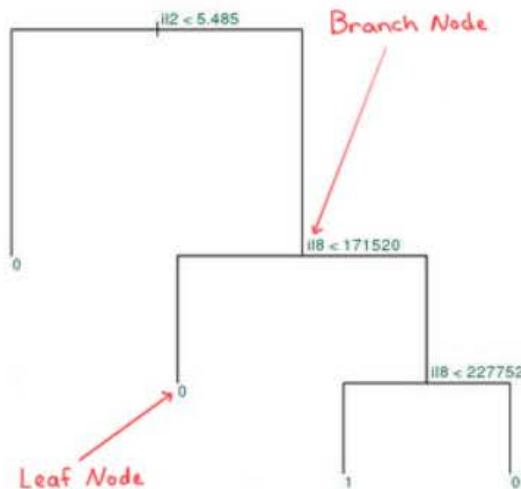
The quality of the model fit is sometimes measured using **accuracy**. Accuracy can be simply defined as the percentage of the subjects in the data set that are correctly labeled 1 for TB or 0 for no TB based on the model. The closer the accuracy is to 100%, the better the model is able to classify who develops TB and who doesn't. 

In our data set, approximately 24% of the subjects developed TB. A model that simply predicted no subjects would develop TB would have an accuracy close to 76%. A model that flipped a coin to decide 0 or 1, would average out to an accuracy of about 50%, because it would be right half the time. A better measure for our purposes is **balanced accuracy**. This is the average of how well the model predicts 1's and how well it predicts 0's. The balanced accuracy for flipping a coin or predicting no TB is the same, 50%, regardless of the percentage of 1's in the dataset. In the rest of this module, when we refer to accuracy, we'll mean balanced accuracy.

An important concern with any data mining exercise is **overfitting**. Overfitting happens when the model includes noisy variables that aren't important. This can result in a high accuracy that has nothing to do with the endpoint. To address overfitting, it is often useful to assess the generalizability of the model. That is, how well does the model do when presented with data it hasn't seen? **Cross-validation** is an approach that is used to divide the data into multiple pieces. A model is fit to one piece of the data and then evaluated on the other piece. The accuracy of the model applied to the second piece is a measure of the generalizability of the model. A model that includes noisy variables that aren't important for the endpoint are not likely to generalize. For more information about cross-validation visit [Wikipedia Cross-validation page](#).

We have divided the TB data into two pieces. Our model will be fit to 2/3 of the data and then evaluated for its accuracy on the remaining 1/3 of the data.

Figure 1



Decision trees (also called **classification trees**) are one of the easiest to understand machine learning methods. The model we discussed previously can be expressed as a decision tree (see Figure 1).

IF $i12 \geq 5.485$ AND $i18 \geq 171520$ AND $i18 < 227752$ THEN $TB=1$ ELSE $TB=0$

Each **branch node** of the tree represents a split of the subjects depending on the value of a single variable. In the first branch node, all the subjects with $i12 < 5.485$ go on the left side of the tree, and the rest go on the right. Each subsequent node subdivides the subjects. Each variable and threshold value is chosen to maximize the separation of $TB=0$ and $TB=1$ (the predicted endpoints). Splitting stops when there are fewer than 10 subjects in a branch or when no variable can produce enough separation of 1's and 0's. These terminal nodes or **leaf nodes** of the tree are labeled 1 if more subjects in that subgroup have $TB=1$ than $TB=0$. This represents the **classification** of that leaf, and allows the model to be used as a predictor on new data.

This model is evaluated by calculating the accuracy across all of the leaves when applied to testing data.

The goal of the computer algorithm that builds the models is to find the right number and combination of variables for the tree (i.e. variable selection) along with the right functions or splits that produce the fewest errors in the leaves.

Your goal in the following exercise is to do the variable selection. We will let the computer do the rest. For more information about decision trees visit: [Wikipedia Decision tree page](#).

← Decision Tree Models

Decision Tree Analysis

Behind The Scenes ⇒

Select five variables. The decision tree algorithm will then find the best fit model for classifying subjects as developing TB using these variables. Note that the algorithm may exclude some of your variables if it determines they aren't useful. Repeat with different combinations to find a high accuracy.

Cytokines

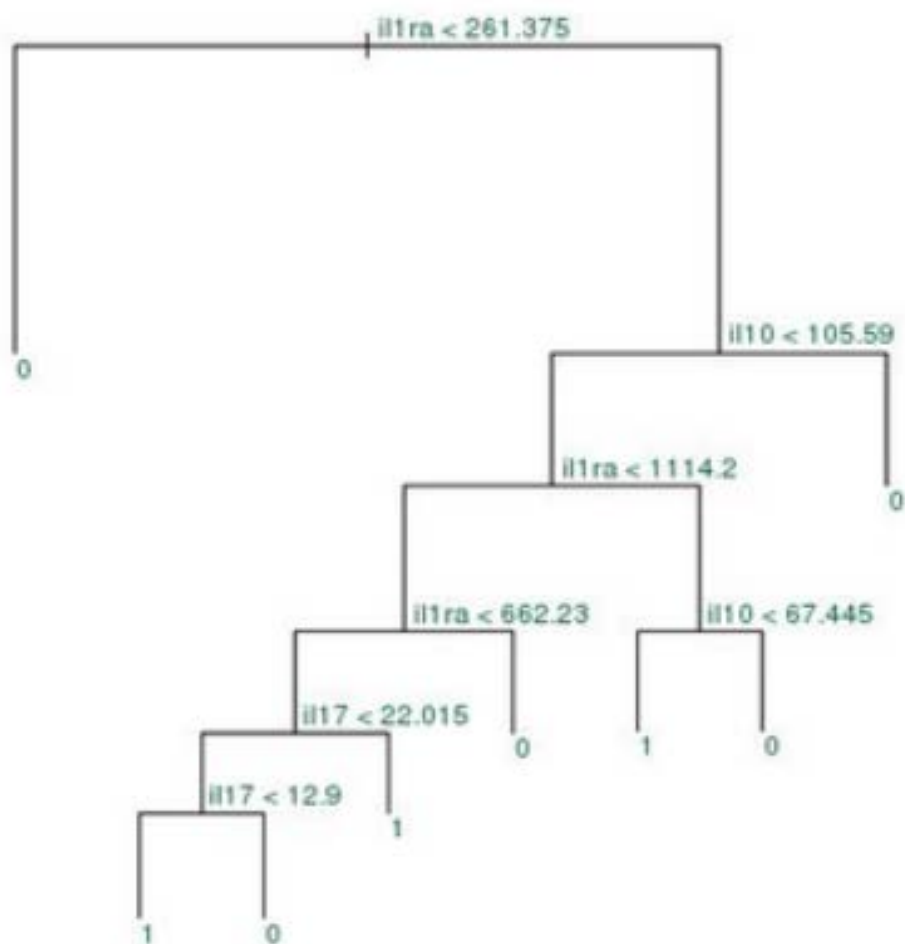
- | | | | | | | | |
|---|--|--|----------------------------------|-----------------------------------|-------------------------------|--------------------------------|---------------------------------|
| <input type="checkbox"/> ccl2 | <input type="checkbox"/> ccl3 | <input type="checkbox"/> ccl4 | <input type="checkbox"/> eotaxin | <input type="checkbox"/> fgfbasic | <input type="checkbox"/> gcsf | <input type="checkbox"/> ifng | <input type="checkbox"/> il1b |
| <input checked="" type="checkbox"/> il1ra | <input type="checkbox"/> il2 | <input type="checkbox"/> il4 | <input type="checkbox"/> il5 | <input type="checkbox"/> il6 | <input type="checkbox"/> il7 | <input type="checkbox"/> il8 | <input type="checkbox"/> il9 |
| <input checked="" type="checkbox"/> il10 | <input type="checkbox"/> il12p70 | <input checked="" type="checkbox"/> il13 | <input type="checkbox"/> il15 | <input type="checkbox"/> il17 | <input type="checkbox"/> ip10 | <input type="checkbox"/> pdgfb | <input type="checkbox"/> rantes |
| <input type="checkbox"/> tnfa | <input checked="" type="checkbox"/> vegf | | | | | | |

Covariates

- | | | |
|----------------------------------|--|------------------------------|
| <input type="checkbox"/> vaccine | <input checked="" type="checkbox"/> gender | <input type="checkbox"/> age |
|----------------------------------|--|------------------------------|

Perform Analysis

Clear Checkboxes

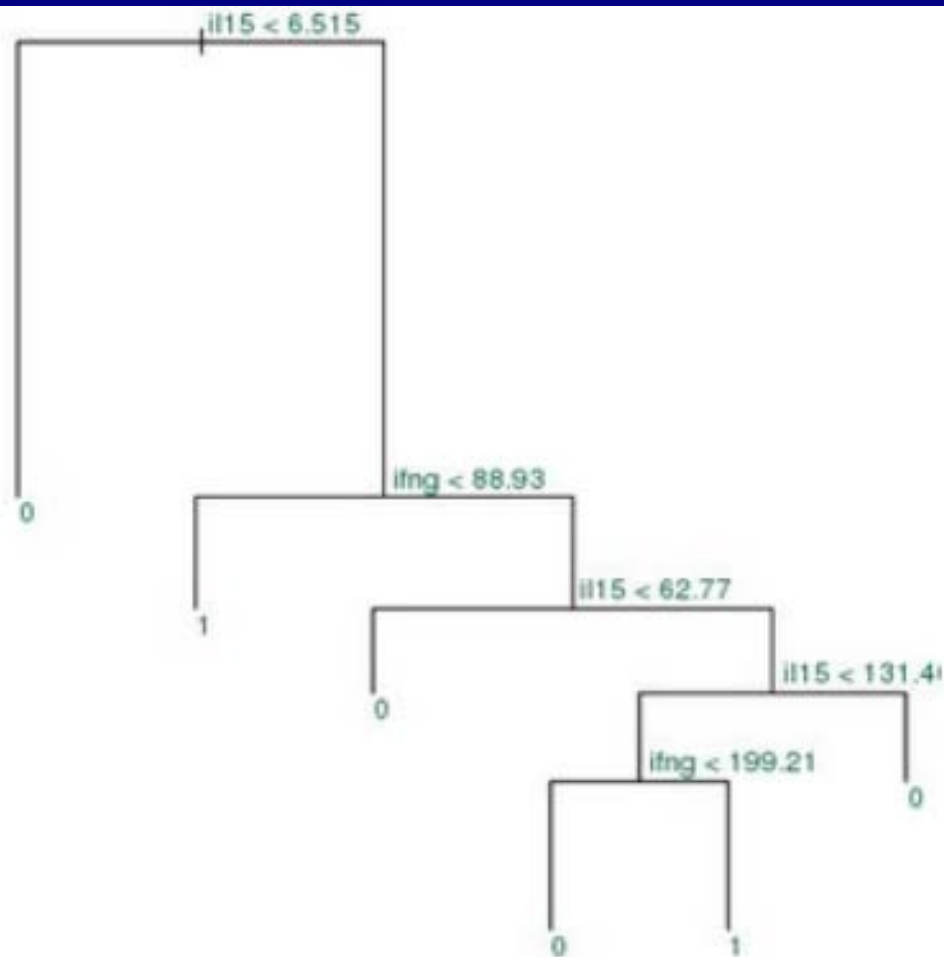


Summary of current results

model	variables					accuracy
1	il1ra	il10	il17	il13	vegf	0.61

Summary of all results (used/excluded variables)

model	variables					accuracy
1	il1ra	il10	il17	il13	vegf	0.61



Summary of current results

model	variables					accuracy
4	il15	ifng	il1ra	il4	vegf	0.67

Summary of all results (used/excluded variables)

model	variables					accuracy
4	il15	ifng	il1ra	il4	vegf	0.67
1	il1ra	il10	il17	il13	vegf	0.61
2	il1ra	il10	il4	il8	vegf	0.541

Data Mining

Grid Computing

Network Science and Modeling

Cluster Analysis

DNA Methylation Array Analysis

Introduction to Statistics

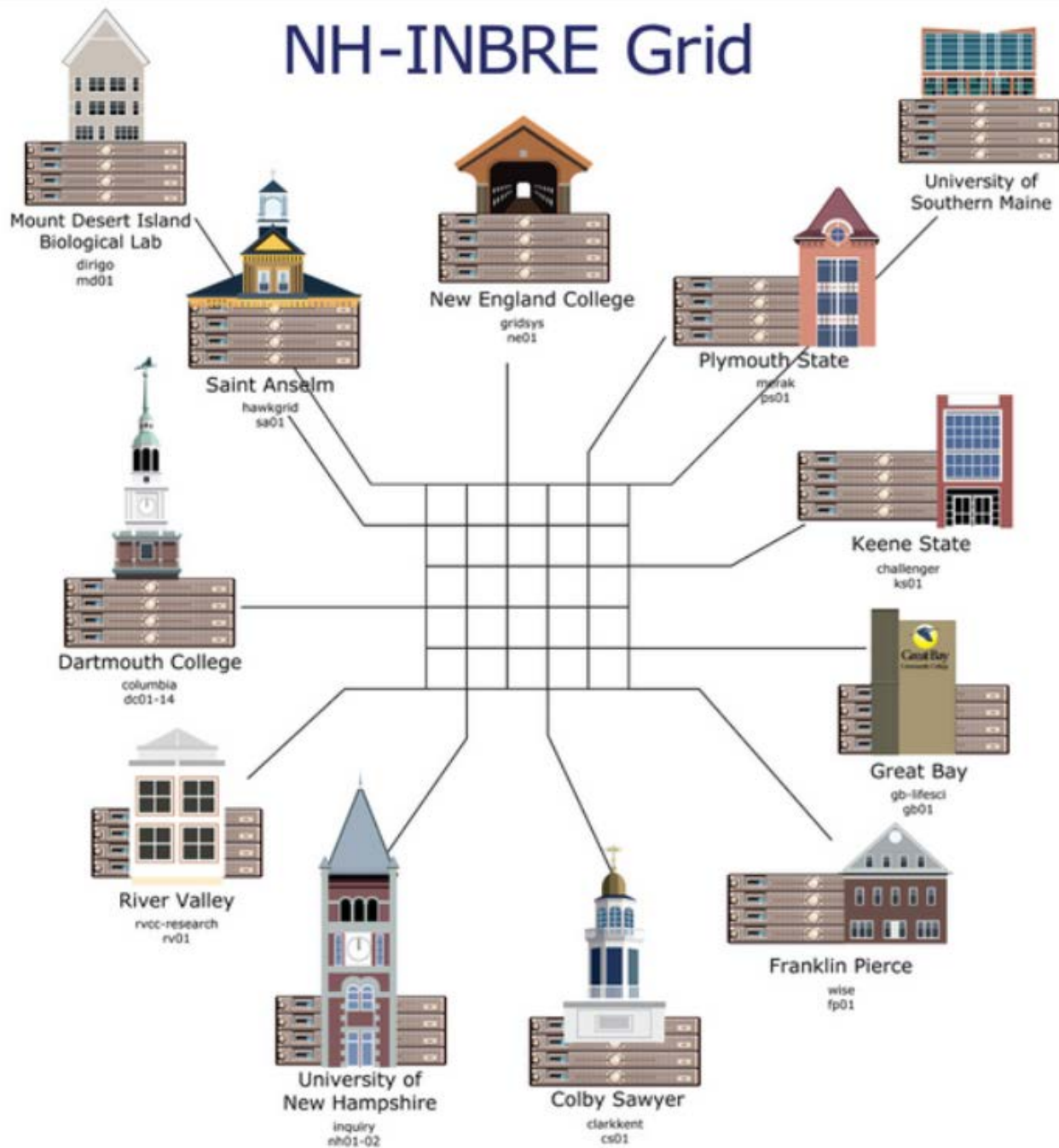
Random Boolean Networks

Hypothesis Testing

Quantitative Real-time PCR

Heatmaps

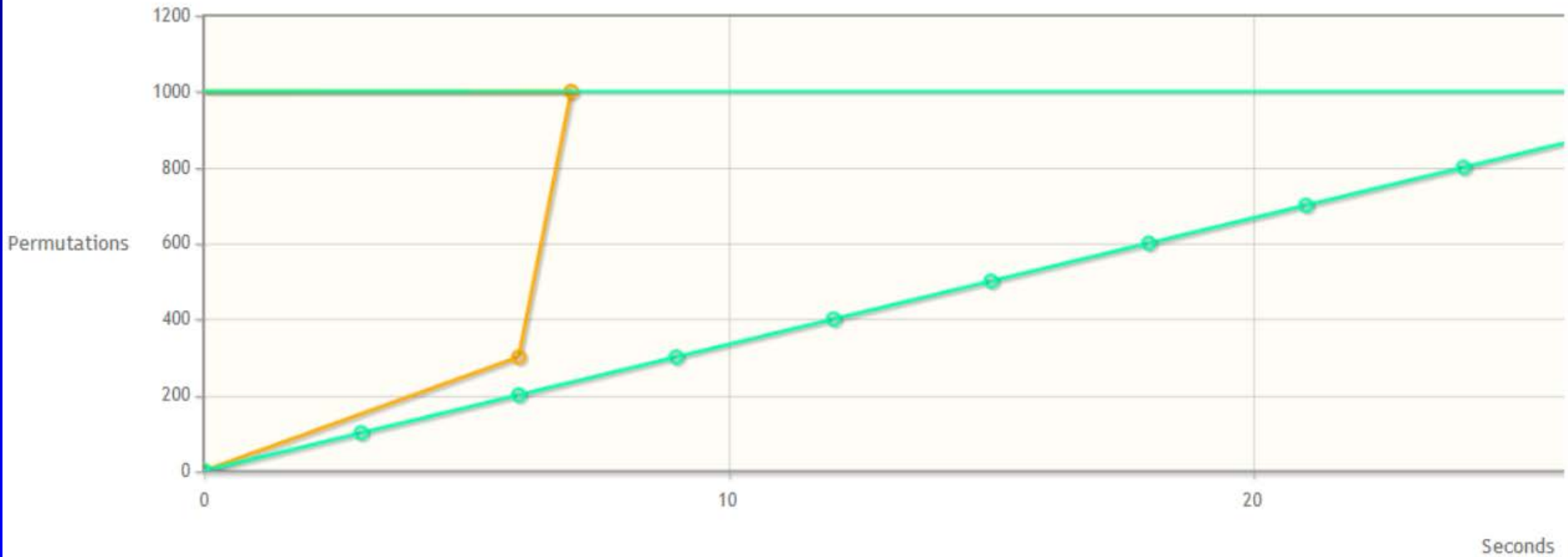
NH-INBRE Grid



Things to note as you run tests: Which clusters take longer? A cluster does a batch of jobs, then another batch. Does batch size relate to number of cores? Does running sequentially beat parallel for this application? Is distributing on the grid better than designating a cluster?

Permutations to run (x100): How to run:

Run Tests 2000 Permutation tests complete

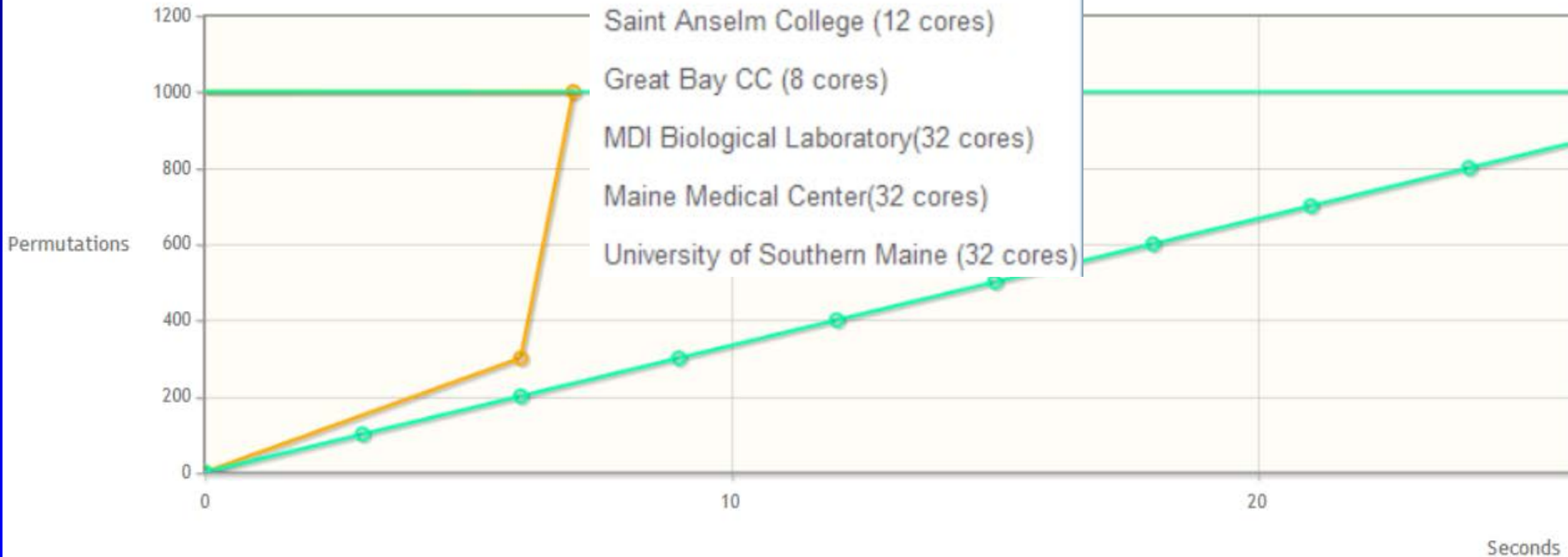


# Requested	Location	# Completed	Mean Accuracy	Execution Seconds	Tests/Second
1000	Dartmouth College (180 cores)	1000	0.499	7	142
1000	Serially	1000	0.499	30	33

Things to note as you run tests: Which clusters take longer
Does running sequentially beat parallel for this application

Permutations to run (x100): How to run:

2000 Permutation tests complete



- Distribute on Grid
- Dartmouth College (180 cores)
 - Colby Sawyer (12 cores)
 - Keene State College (12 cores)
 - Plymouth State University (12 cores)
 - Saint Anselm College (12 cores)
 - Great Bay CC (8 cores)
 - MDI Biological Laboratory(32 cores)
 - Maine Medical Center(32 cores)
 - University of Southern Maine (32 cores)

ch. Does batch size relate to number of cores?
a cluster?

# Requested	Location	# Completed	Mean Accuracy	Execution Seconds	Tests/Second
1000	Dartmouth College (180 cores)	1000	0.499	7	142
1000	Serially	1000	0.499	30	33

Acknowledgments

- INBRE
- Doug Hill
- Katrina Bogan
- Christian Darabos
- Anne Hoen
- Jing Li
- Ryan Urbanowicz