Simplifying Drug Discovery with JMP

John A. Wass, Ph.D.
Quantum Cat Consultants, Lake Forest, IL

Cele Abad-Zapatero, Ph.D.
Adjunct Professor, Center for Pharmaceutical Biotechnology, University of Illinois at Chicago, Chicago, IL

Introduction

Pharmaceutical product development consists of a research phase, called Drug Discovery, and an applied phase where clinical studies are performed. In both areas, sophisticated and expensive software is extensively utilized to provide analytic power and database management. In the research phase, there are very many steps where simpler methods and less costly software may be utilized for a variety of studies. Although academe, where much of the research is done, favors such freeware as R and various 'roll-your-own' software, industry and government often use commercial grade software such as JMP® where powerful analytics are paired with sophisticated graphics in easy to use format. This paper delves into two areas of Discovery, gene expression, and ligand efficiency indices as they relate to mapping Chemico-biological space.

Methods I: Gene Expression

Genomic Datasets tend to be large and need to be statistically treated to properly convey the true nature of the underlying physical processes. These treatments include numerous issues of statistical normalization and multi-sample testing as well as biochemical variability, probe-sample pairing, and gene nomenclature which are not the subject of this brief presentation. As we wish to concentrate on the actual multivariate analysis of the data, we begin after the initial step of identifying the minimal gene set thought to be responsible, or at least contributory to the effect under study.

In this section, we employ the Multivariate platform in JMP®8 to analyze tumor cell line resistance to analyze tumor cell line resistance to a metabolic inhibitor (1,2). Specifically this platform is used to i) separate resistance classes by Discriminant analysis and Clustering, and ii) further examine these differences by Principal Components Analysis (PCA). It further employs the Partition and Fit Model techniques to verify results. These types of strategies have been employed in many instances in drug discovery and find widespread use in molecular biology. Studies such as these are now implemented and greatly extended in JMP® Genomics 5.0.

It is the purpose of this paper to show how straightforward multivariate methods may shed light on the relative importance of various genes and gene groups. The data are from a microarray study and represent fluorescence intensities from a microarray reader. The study compared expression of a family of pro-apoptotic genes and sensitivity to an anti-tumor drug in a panel of small cell lung carcinoma (SCLC), leukemia, and lymphoma cell lines, grouped as sensitive, intermediate, or resistant to anti-tumor drugs. The analytic methods here, from the Multivariate Methods, Modeling, and Fit Model Platforms are:

Clustering (a multivariate grouping technique that joins groups having similar properties. Ward's method is used for small tables of less than a few thousand rows. Numeric variables are utilized where numeric differences make sense physically)

Discriminant Analysis (useful to group variables by predicting class membership). The class variables may be nominal or ordinal but the predictor variables must be continuous. JMP uses a common covariance matrix for all groups and calculates the distance between points and their multivariate means.

PCA (employs linear combinations of the standardized variables to reduce dimensionality of complex data sets and separate groups on the basis of variance vectors)

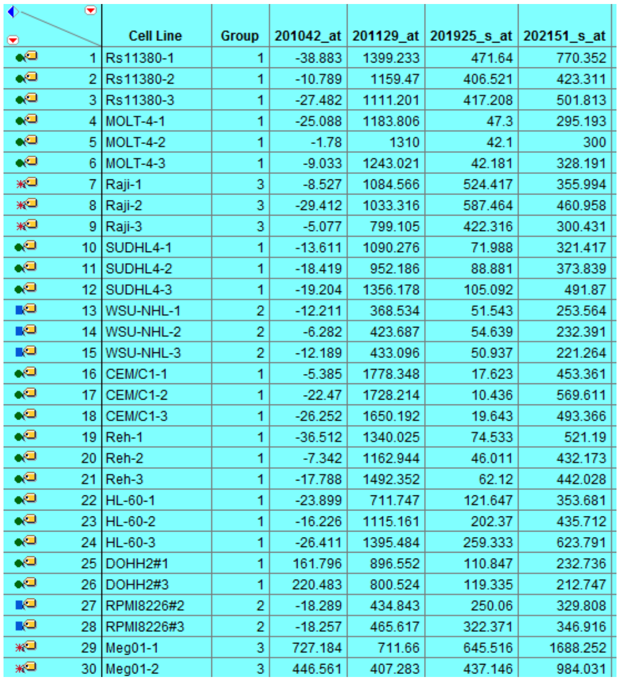
Partition Analysis (partitions data according to a relationship between the X and Y values and determine groupings of X values that best predict a Y value)

Fit Ordinal Logistic Model (the Fit Model platform utilizes the ordinal logistic fitting personality to model non-continuous data)

These methods are used to i) asses their ability to cleanly separate the three groups (levels) of drug resistance and ii) be used as a model to predict resistance in new drugs.

Data

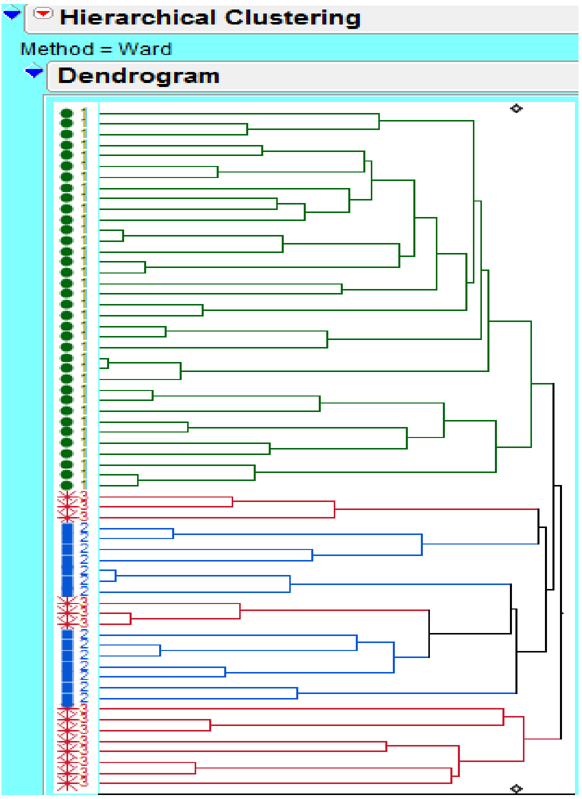
A portion of the data set is displayed below:



		Cell Line	Group	201042_at	201129_at	201925_s_at	202151_s_at
●	1	Rs11380-1	1	-38.883	1399.233	471.64	770.352
●	2	Rs11380-2	1	-10.789	1159.47	406.521	423.311
●	3	Rs11380-3	1	-27.482	1111.201	417.208	501.813
●	4	MOLT-4-1	1	-25.088	1183.806	47.3	295.193
●	5	MOLT-4-2	1	-1.78	1310	42.1	300
●	6	MOLT-4-3	1	-9.033	1243.021	42.181	328.191
✳	7	Raji-1	3	-8.527	1084.566	524.417	355.994
✳	8	Raji-2	3	-29.412	1033.316	587.464	460.958
✳	9	Raji-3	3	-5.077	799.105	422.316	300.431
●	10	SUDHL4-1	1	-13.611	1090.276	71.988	321.417
●	11	SUDHL4-2	1	-18.419	952.186	88.881	373.839
●	12	SUDHL4-3	1	-19.204	1356.178	105.092	491.87
■	13	WSU-NHL-1	2	-12.211	368.534	51.543	253.564
■	14	WSU-NHL-2	2	-6.282	423.687	54.639	232.391
■	15	WSU-NHL-3	2	-12.189	433.096	50.937	221.264
●	16	CEM/C1-1	1	-5.385	1778.348	17.623	453.361
●	17	CEM/C1-2	1	-22.47	1728.214	10.436	569.611
●	18	CEM/C1-3	1	-26.252	1650.192	19.643	493.366
●	19	Reh-1	1	-36.512	1340.025	74.533	521.19
●	20	Reh-2	1	-7.342	1162.944	46.011	432.173
●	21	Reh-3	1	-17.788	1492.352	62.12	442.028
●	22	HL-60-1	1	-23.899	711.747	121.647	353.681
●	23	HL-60-2	1	-16.226	1115.161	202.37	435.712
●	24	HL-60-3	1	-26.411	1395.484	259.333	623.791
●	25	DOHH2#1	1	161.796	896.552	110.847	232.736
●	26	DOHH2#3	1	220.483	800.524	119.335	212.747
■	27	RPMI8226#2	2	-18.289	434.843	250.06	329.808
■	28	RPMI8226#3	2	-18.257	465.617	322.371	346.916
✳	29	Meg01-1	3	727.184	711.66	645.516	1688.252
✳	30	Meg01-2	3	446.561	407.283	437.146	984.031

The cell lines are identified in the first column, the sensitivity class in the second column and all others are the gene designations for the minimal gene set identified by ANOVA.

To see if we can really (cleanly) separate the drug sensitivity classes by the gene expression changes, a first step is to cluster the data. This is a rough-cut technique and depends upon the technique utilized. We choose Hierarchical clustering as it is appropriate to small samples and we choose Ward’s method as the best general compromise for a preliminary analysis. The results displayed below, indicate that one group (sensitive) is completely separated, while the other two (moderately sensitive and resistant) are not.

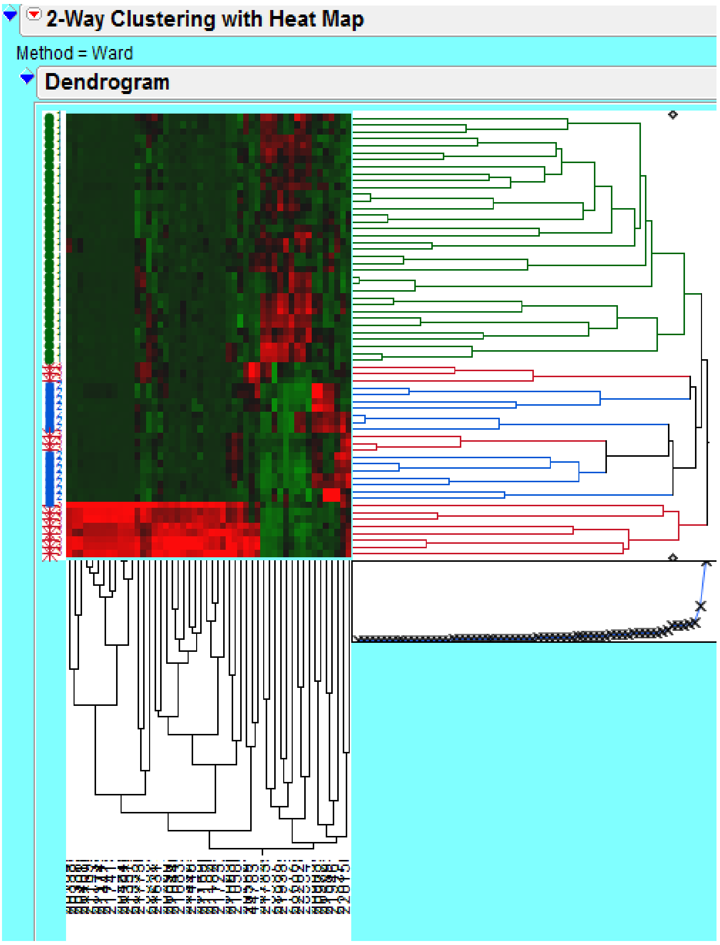


Hierarchical Clustering

Method = Ward

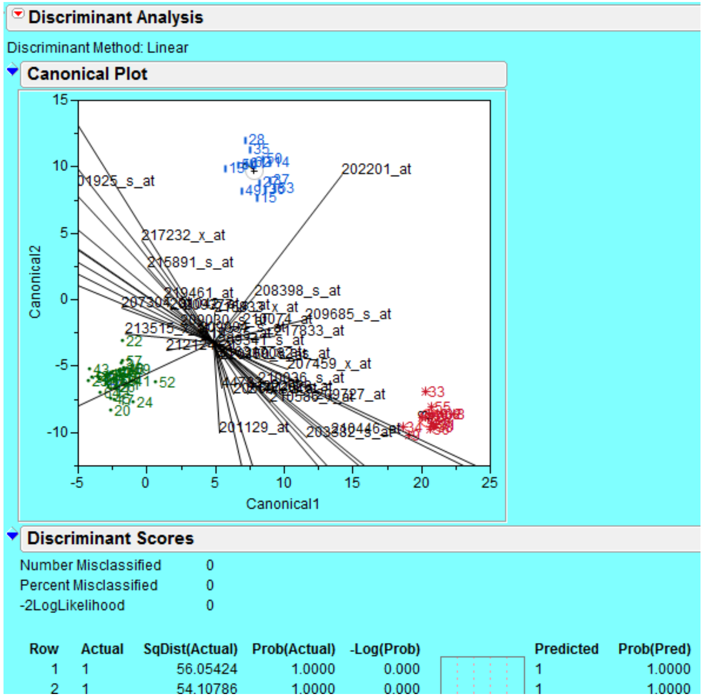
Dendrogram

The list can be further clustered in 2D showing a heat map and separation both by sensitivity group and genes. This type of clustering might uncover important similarities in pathway action by the different genes.



Method = Ward

Another way to approach the class separation problem is Discriminant Analysis and will usually make a cleaner cut. DA will ‘measure the distance from each point in the data set to each group's multivariate mean (often called a centroid) and classifies the point to the closest group. The distance measure used is the Mahalanobis distance, which takes into account the variances and covariances between the variables’ (3). For this separation a linear DA was used where ‘it is assumed that the Y's are normally distributed with the same variances and covariances, but that there are different means for each group defined by X’ (3).



Discriminant Analysis

Discriminant Method: Linear

Canonical Plot

Discriminant Scores

Number Misclassified 0
Percent Misclassified 0
-2LogLikelihood 0

Row	Actual	SqDist(Actual)	Prob(Actual)	-Log(Prob)		Predicted	Prob(Pred)
1	1	56.05424	1.0000	0.000		1	1.0000
2	1	54.10786	1.0000	0.000		1	1.0000

The finest cut is usually made with Principal Components Analysis. This highly mathematical method will use as many dimensions as necessary to effect a clean separation. Luckily, in most cases, only 2 or 3 are needed so the results may be interpreted physically. PCA is very useful to derive a small number of independent linear combinations of a set of variables that capture as much of the variability in the original variables as possible.

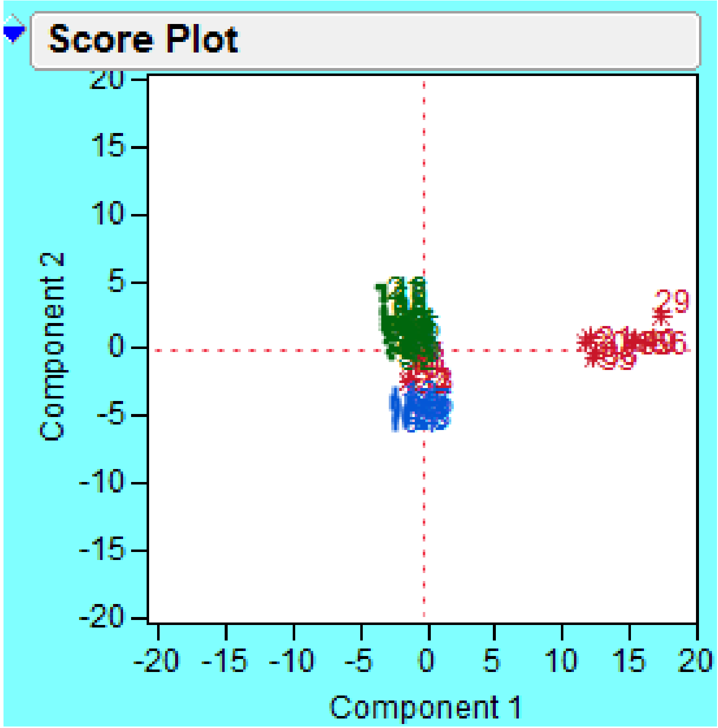
'Using principal component analysis reduces the dimensionality of a set of data. Principal components (PC's) are a way to picture the structure of the data as completely as possible by using as few variables as possible.

For n original variables, n principal components are formed as follows:

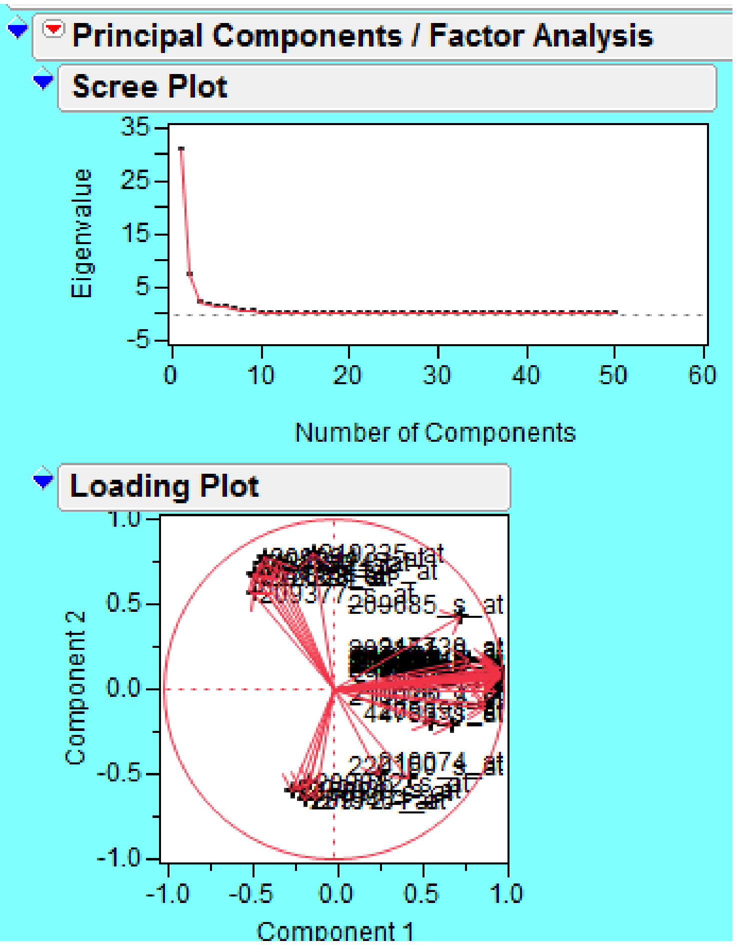
- The first principal component is the linear combination of the standardized original variables that has the greatest possible variance.
- Each subsequent principal component is the linear combination of the standardized original variables that has the greatest possible variance and is uncorrelated with all previously defined components.

Each principal component is calculated by taking a linear combination of an eigenvector of the correlation matrix with a standardized original variable. The eigenvalues show the variance of each component' (3).

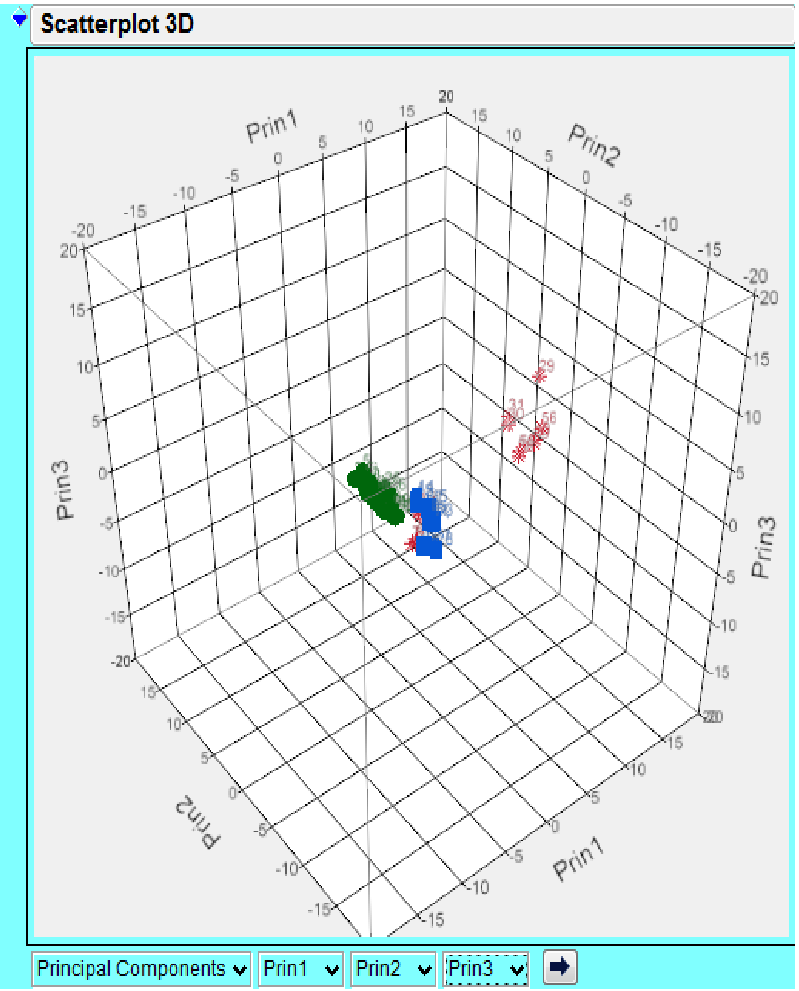
In JMP we can easily visualize the group separations with a score plot, that shows how the first two principal components are used to separate them



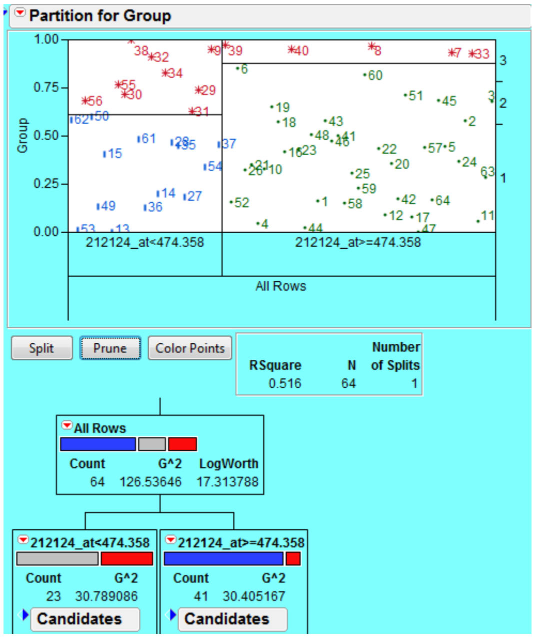
At the same time, JMP produces diagnostics that let us know that two PC's are sufficient for the separation (Scree Plot) and the relative importance of the individual genes in forming the vectors (Loading Plot).



We can also produce a rotateable 3D scatterplot to better visualize the separation. This is important when placing new, unknown drugs into the calculations.



Finally, Partition Analysis and Logistic Regression may be used as alternative ways to i) separate the classes and ii) gain a deeper understanding of which genes may be driving the separation and thus drug sensitivity or resistance. Recursive partitioning is especially useful for preliminary exploration when there is a large data set with no prior model. The results are straightforward to interpret. Logistic regression is useful for fitting non-continuous x-axis (predictor) data and can be effective with a single predictive factor.

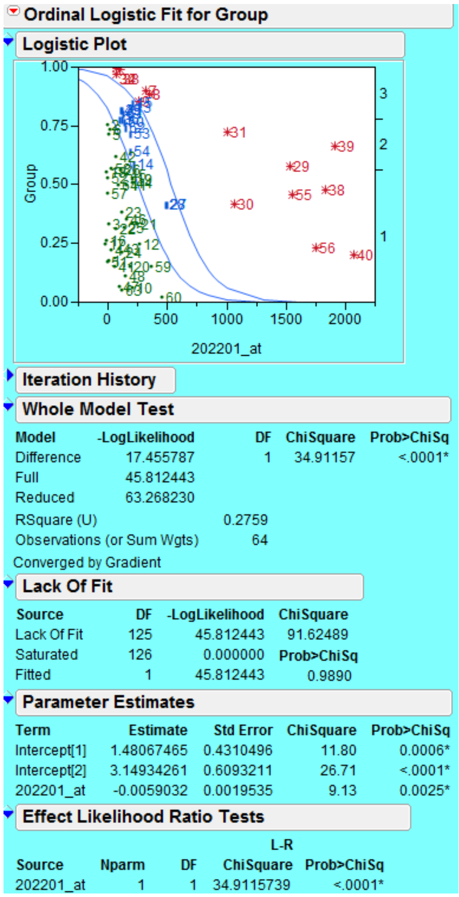


```
RSquare      N   Number
                 of Splits
  0.516     64          1
```

```
All Rows
Count      G^2      LogWorth
   64   126.53646   17.313788
```

```
212124_at<474.358
Count      G^2
   23   30.789086
Candidates
```

```
212124_at>=474.358
Count      G^2
   41   30.405167
Candidates
```



```
Whole Model Test
Model          -LogLikelihood        DF   ChiSquare   Prob>ChiSq
Difference         17.455787          1    34.91157      <.0001*
Full               45.812443
Reduced            63.268230

RSquare (U)                     0.2759
Observations (or Sum Wgts)          64
Converged by Gradient
```

```
Lack Of Fit
Source        DF   -LogLikelihood   ChiSquare
Lack Of Fit   125       45.812443    91.62489
Saturated     126        0.000000   Prob>ChiSq
Fitted          1       45.812443      0.9890
```

```
Parameter Estimates
Term            Estimate    Std Error   ChiSquare   Prob>ChiSq
Intercept[1]   1.48067465   0.4310496       11.80      0.0006*
Intercept[2]   3.14934261   0.6093211       26.71      <.0001*
202201_at     -0.0059032    0.0019535        9.13      0.0025*
```

```
Effect Likelihood Ratio Tests
                                 L-R
Source      Nparm   DF    ChiSquare   Prob>ChiSq
202201_at       1    1   34.9115739      <.0001*
```

Methods II: Ligand Efficiency Indices (Mapping of Chemo-Biological Space)

Yet another approach now being developed in drug discovery is the use of Ligand Efficiency Iindices (LEI's) to produce maps of chemical/biological properties of compounds, drug fragments and actual approved drugs. When new fragments (or chemical entities) are placed upon the map they may indicate the new fragments 'similarity' (i.e., proximity) to an approved, on-market drug and thus suggest the probability that this fragment may successfully complete clinical trials. Although the chemistry runs deep and requires a multidimensional optimization of chemical structure integrating enzymology and cellular and organismic biology, the mathematics behind it is straightforward and the results easy to understand, interpret, and implement by the medicinal chemist (4).

Historically it was realized that a more streamlined and efficient process for drug discovery was needed as increasingly, more money and effort is being used to less effect. As Lipinski's "Rule of Five" is limited in several ways and becoming less relevant to the discovery process, attention was focused on the potency of ligands to receptors (e.g., Ki) and evaluation of the fragments properties such as molecular weight, solubility, polar surface area (PSA), and ClogP. Beyond simple filtering there was a need to construct a numerical framework capable of predictive ability to maximize the probability of identifying a viable drug candidate.

Ligand efficiency (LE) was originally defined in terms of ΔG and the number of non-hydrogen atoms N in the compound (5): LE = ΔG/N, where ΔG is the familiar Gibbs free energy and defined as $\Delta G = -RT\ln K_{i.}$ Utilizing the importance of molecular weight (MW), the concept was extended to the use of binding affinities (4). As initial calculations demonstrated the efficacy of directly relating the potency (Ki, IC50, or related measurements) to the MW (in kDa) and PSA of the ligands, these became the basis of the development of maps of the chemo-biological space that provide enhanced understanding of the distribution of compounds in terms of PSA/MW across the polarity-efficiency vs size-efficiency (SEI-BEI) plane (4,6). Moreover this representation suggests that compounds with both large values of SEI and BEI, optimize both variables simultaneously (6). In time the concepts were expanded to include related factors such as the number of heavy (non-Hydrogen) atoms and the number of polar atoms (i.e., O and N).

When we do this, we derive some very useful tools. Basically, the maps are constructed on sound chemical principles to display the nBEI-NSEI space (7) that nicely display the fragments and compounds along lines of distinct slope defined by the number of polar atoms in the chemical entity. These maps then, with descriptive annotation, suggest the probabilities of a fragment becoming a successful drug as successful drugs placed on the map suggest that there is i) an area on the lower left where there is little probability that a molecular entity will become a drug and ii) the further out and up in the lines that a new candidate lies, the greater is the probability of successful drug development. A more rigorous analysis is being undertaken to estimate probabilities for different targets.

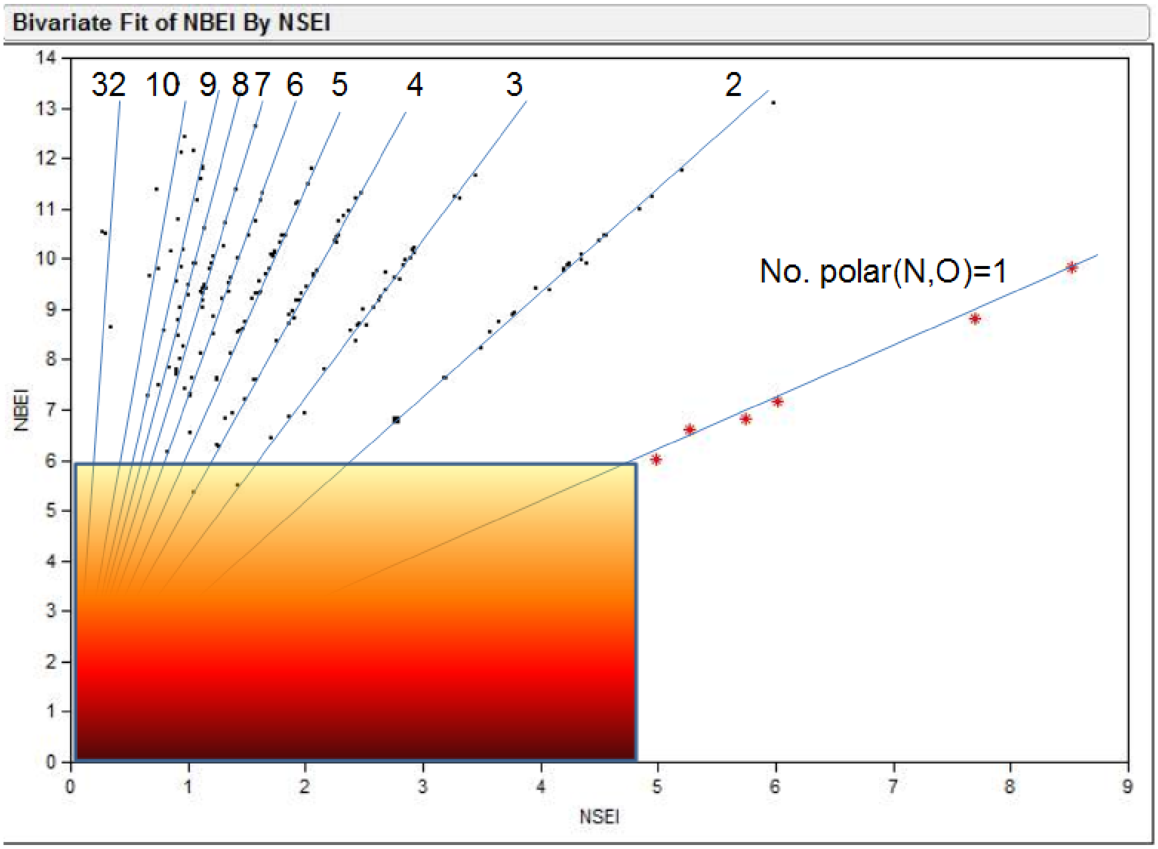
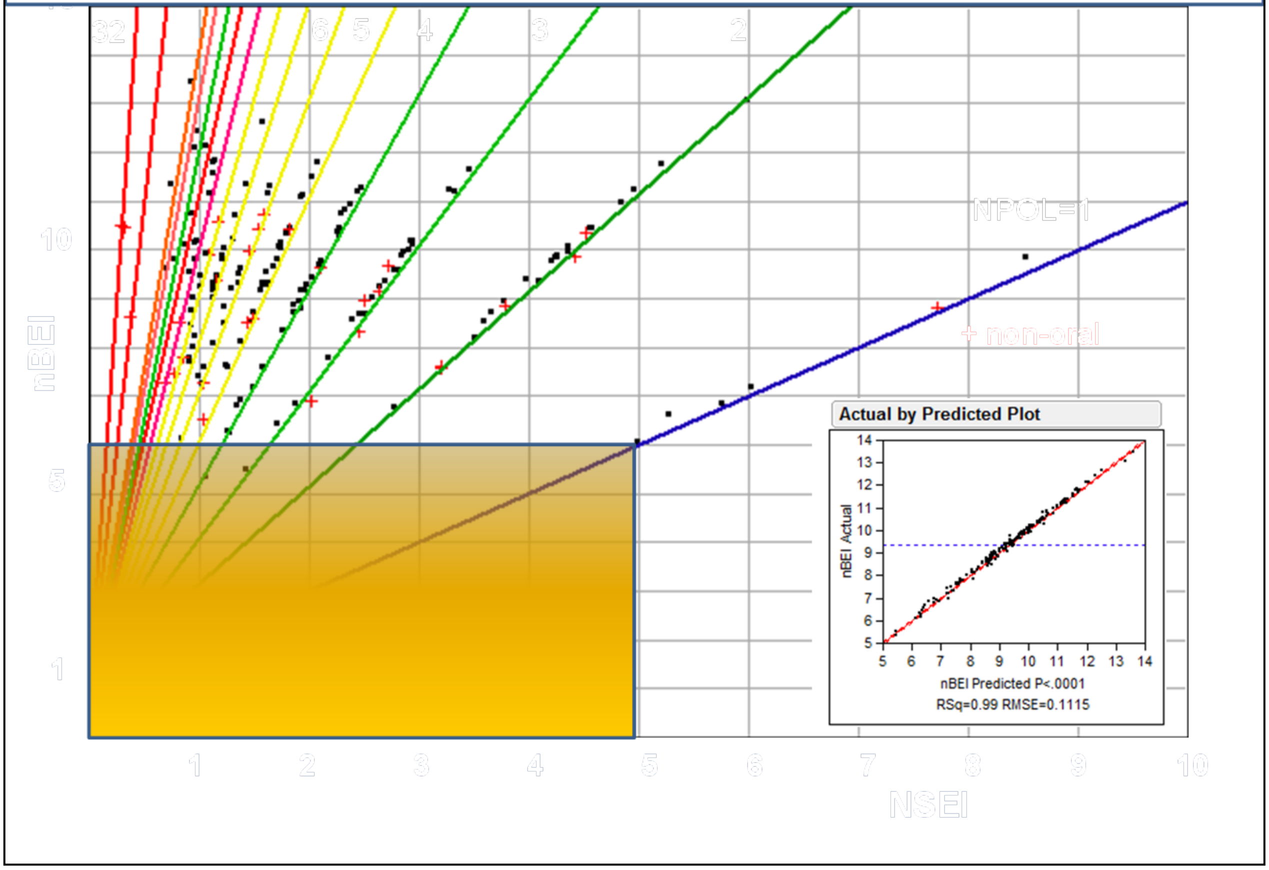
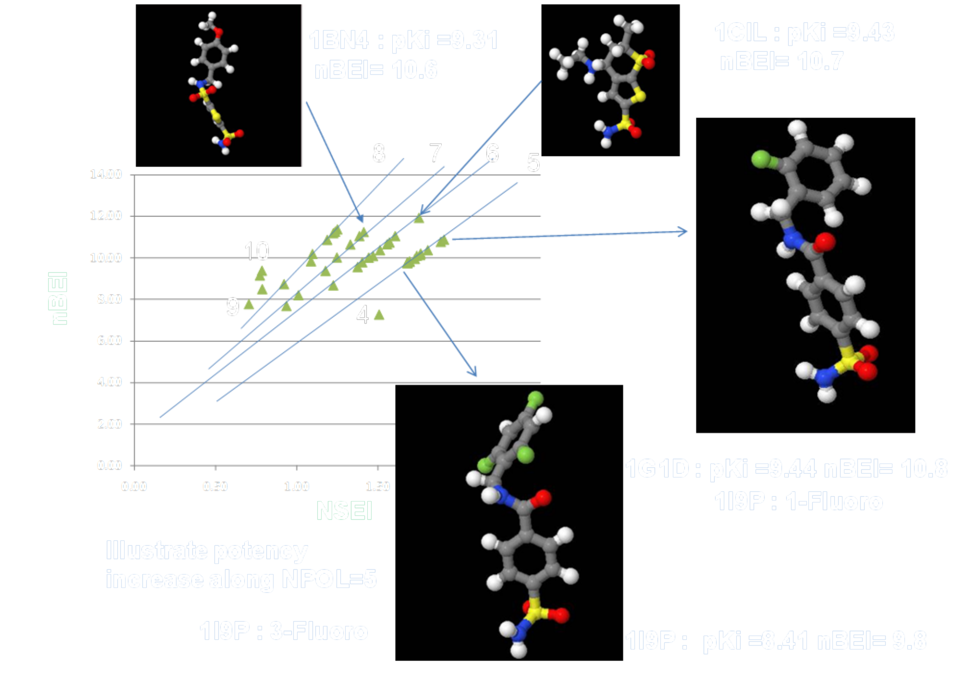


Fig. 1.
Y-axis: nBEI ; X-axis: NSEI. Representation of a limited sample of 200 marketed drugs in the nBEI-NSEI plane. The slope of the lines is equal to the number of polar atoms (N plus O) in the chemical compound. The limited dataset courtesy of the ChEMBL database (European Bioinformatics Institute, Hinxton, Wellcome Trust Campus, Cambridge, UK). See references (4, 6-7) for details.

The "fan" plot for a sample of marketed drugs



Monitoring SAR for different scaffolds along lines of NPOL



1BN4 : pKi =9.31 nBEI= 10.6

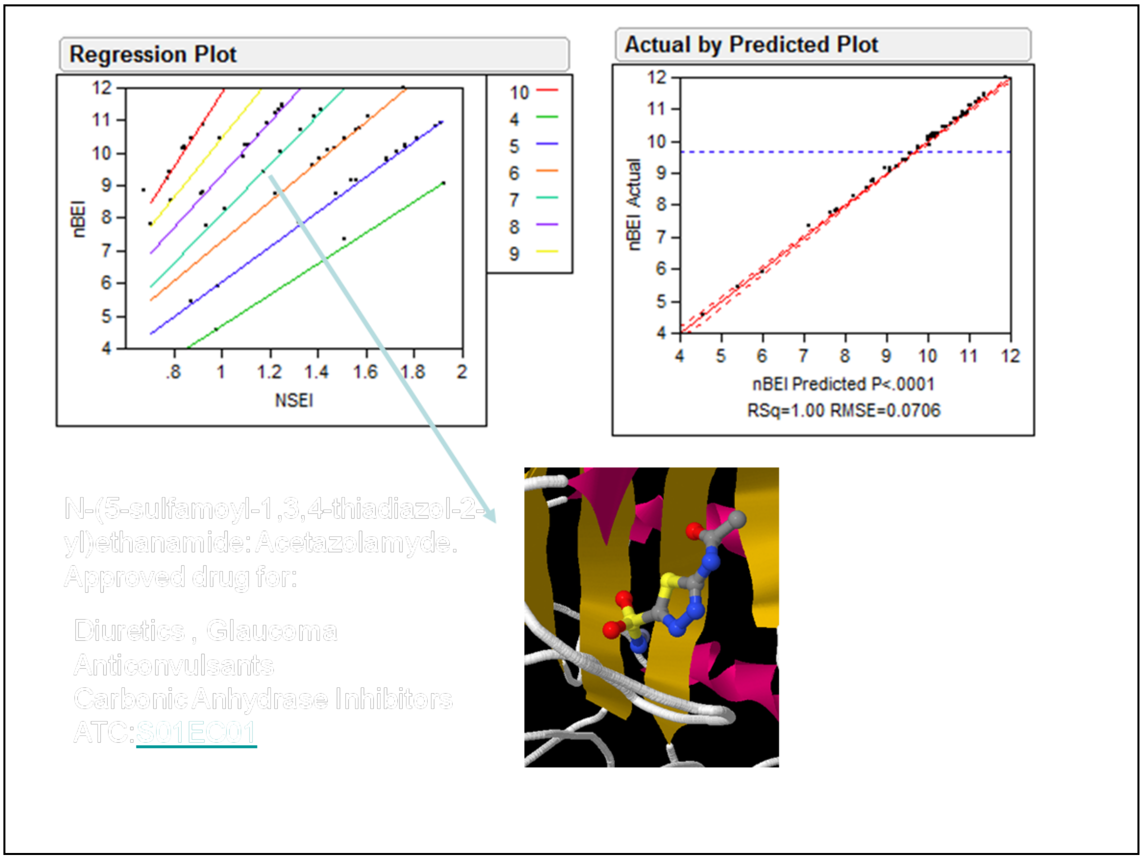
1CIL : pKi =9.43 nBEI= 10.7

1G1D : pKi =9.44 nBEI= 10.8
1I9P : 1-Fluoro

Illustrate potency increase along NPOL=5

1I9P : 3-Fluoro

1I9P : pKi =8.41 nBEI= 9.8



N-(5-sulfamoyl-1,3,4-thiadiazol-2-yl)ethanamide: Acetazolamyde.
Approved drug for:

Diuretics , Glaucoma
Anticonvulsants
Carbonic Anhydrase Inhibitors
ATC:S01EC01

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

References

1. Zeru, M. et. al. p53 and BCl-2 protein expression and its relationship with prognosis in small-cell lung cancer. Clin Lung Cancer 2003; 4:298-302.

2. Tahir S.K. et. al. Identification of expression signatures predictive of sensitivity to the BCl-2 family member inhibitor ABT-263 in small cell lung carcinoma and leukemia/lymphoma cell lines. Mol Cancer Ther. 2010 Mar;9(3):545-57.

3. SAS Institute Inc. 2008. JMP® 8 Statistics and Graphics Guide, Volumes 1 and 2. Cary, NC

4. Abad-Zapatero, C. and J. T. Metz. Ligand efficiency indices as guideposts for drug discovery. Drug Discovery Today (2005) 10(7): 464-469.

5. Hopkins, AL, Groom, CR, and A. Alex. Ligand efficiency: a useful metric for lead selection. Drug Discovery Today (2004) 9(10): 430-431.

6. Abad-Zapatero, C. Ligand Efficiency Indices for Effective Drug Discovery. Expert Opinion in Drug Discovery (2007) 2(4): 469-488.

7. Abad-Zapatero et. al. Ligand Efficiency Indices for an Effective Mapping of Chemico-Biological Space. Drug Discovery Today, 2010. (in press).