Providing Cheminformatics Solutions To Support Drug Discovery Decisions

CARLETON R. SAGE, KEVIN R. HOLME, NIANISH SUD AND RUDY POTENZONE

Lionbioscience, Inc., 9880 Campus Point Dr., San Diego, CA 92121, USA

E-Mail: rudolph.potenzone@lionbioscience.com

Received: 1st August 2002 /Published: 15th May 2003

Abstract

Drug discovery programs have had to deal with an avalanche of data coming from both the adoption of new technologies such as high throughput screening and combinatorial chemistry, as well as advances in genomics and structural genomics which have facilitated a gene family target approach to drug discovery. Although this data rich environment has been a challenge to manage, it has provided an opportunity for the development of informatics based tools and solutions to extract information from this large body of data, and convert this information into knowledge that can be used and reused for drug discovery.

In the cheminformatics field there has been considerable focus on the development of new tools to visualise and analyse the data, particularly with relation to identifying new leads, and analysing SAR for lead optimisation. While individual cheminformatics tools are critical for analysing this data, a real opportunity exists to provide solutions that synthesize results from these analyses into knowledge to support drug discovery decisions. This remains largely a "manual" activity that takes place within individual project teams.

This paper will describe some concepts and implementations of cheminformatics solutions that begin to address the need for reusable knowledge generation within drug discovery projects. The talk will address requirements for the integration of chemical and biological data as well as the integration of tools and models. The power of using predictive tools for compound design will be highlighted as well as methods to simultaneously consider multiple SAR's. We will describe how providing such solutions

INTRODUCTION

The drug discovery process used to be less complicated. Teams of chemists and biologists (drug discovery scientists) would work together on tens to hundreds of molecules to try to specifically alter the function of their biological target. Things have changed. Because of changes in available technologies and increases in fundamental understandings of biology, the drug discovery scientist has to contend with thousands to millions of molecules interacting with potentially thousands of targets. Therefore, the modern drug discovery scientist is awash in data. However, the changes in available technologies haven't necessarily resulted in improvements in the quality of the data. As a result, we have been flung into what might be a morass of meaningless data, or discovery knowledge nirvana. How do we navigate?

COMPONENTS REQUIRED TO BUILD THE MAP

The first stage to approaching this information overload is by assembling the informatics components necessary for integrating all the available data. Databases are an integral component. This simple component may pose problems for some organisations since the default drug discovery data repository has become Microsoft Excel. Once the data has been arranged into databases, establishing a link between the data in the databases is an essential component. This component does not have a trivial solution, since it involves linking data of different types across different research areas. One obvious potential solution is to use the experimental assay data as the common link between the genomics/bioinformatics/proteomics data and cheminformatics data. Assuming that the data has been integrated to some extent, the next required components are ways to visualise, analyse, and interrogate the data. Furthermore, sophisticated computational approaches must also be taken in order to summarise collections of data into reusable knowledge for prospective application. Finally, means to allow `rollup' of these components for decision support, including tracking and knowledge management tools are essential for efficient. Figure 1 illustrates the Lion Bioscience architecture.

Once an integrated system of data, analysis, and decision support tools has been created, then the true power of the system can be exploited for more rational/informed decision making. These systems can be used in all phases of the drug discovery process from target selection to screening set selection to lead optimisation and lead selection for pre-clinical development.

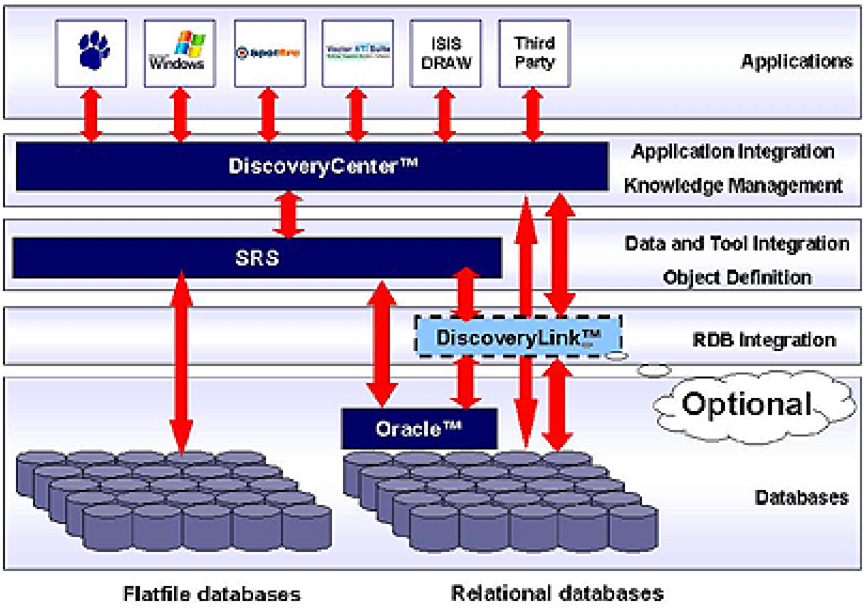
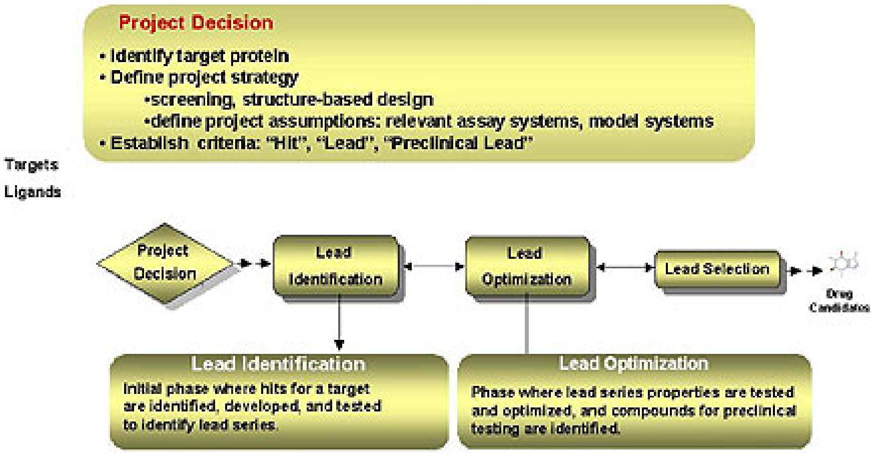


Figure 1. Lion Bioscience integration architecture

The purpose of this paper is to illustrate the potential utility of these approaches in the post-target selection region of drug discovery, and the next two figures illustrate a framework for discussing the small molecule drug discovery process. Figure 2 defines our use of the terms "Lead Identification" and "Lead Optimisation" since these terms may have different connotations in different organisations.



Project Decision

- Identify target protein
- Define project strategy
  - •screening, structure-based design
  - •define project assumptions: relevant assay systems, model systems
- Establish criteria: "Hit", "Lead", "Preclinical Lead"

Initial phase where hits for a target are identified, developed, and tested to identify lead series.

Phase where lead series properties are tested and optimized, and compounds for preclinical testing are identified.

Figure 2. Definitions and Assumptions for Drug Discovery project initiation.

A second concept illustrated in figure 2 are some of the operating assumptions and project criteria that must be performed in order for this integrated approach to work most efficiently. Project criteria are a key decision support / project tracking consideration. Figure 3 illustrates a version of a pre-clinical drug discovery (or lead candidate selection) workflow. It is important to note that this entire workflow sits upon the foundation of knowledge/information gleaned from other projects, and any and all data (both positive and negative) and knowledge generated for a given project in this workflow is put back into this foundational "Knowledge Base".

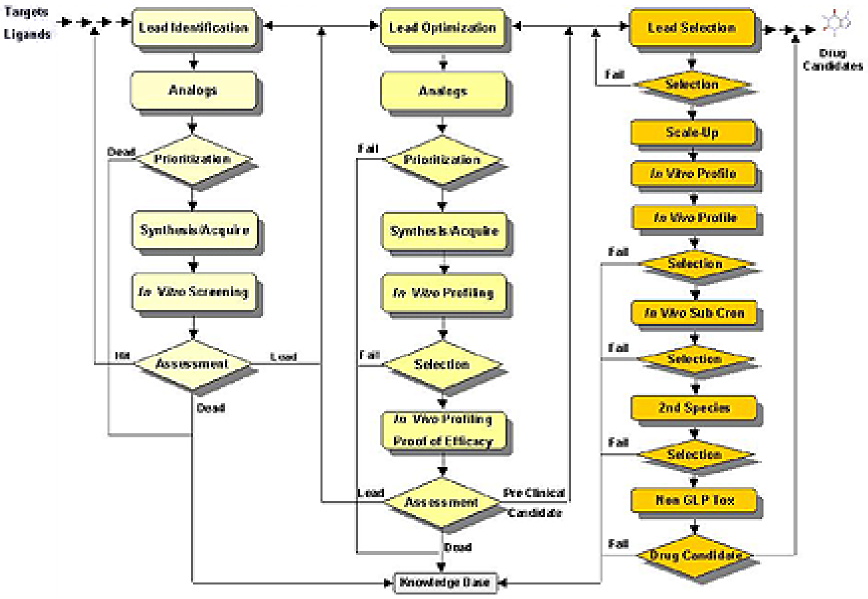


Figure 3. Idealised project work/information flow for Pre-clinical Drug Discovery.

Reusing and Leveraging Existing Data Through Computational Models (Hypotheses)

Leveraging data that exists at project initiation via computational means provides momentum in the rational decision making early in the drug discovery process, and continues this momentum as more data becomes available. The type, quantity, and quality of these data define the extent to which computational approaches should be used. Figure 4 describes the potential computational approaches that could be used in support of drug discovery projects given defined sets of experimental data.

In the extreme, if the project under consideration has a liganded protein structure, related protein structures and lead series data available, computational approaches including diversity, similarity, QSAR/pharmacophore models, drug-likeness models, docking, cross-target (or specificity) alert models, and project-specific ADME-models can be applied simultaneously to best leverage all existing information in the compound prioritisation and compound assessment phases of lead identification and lead optimisation (Figure 3).

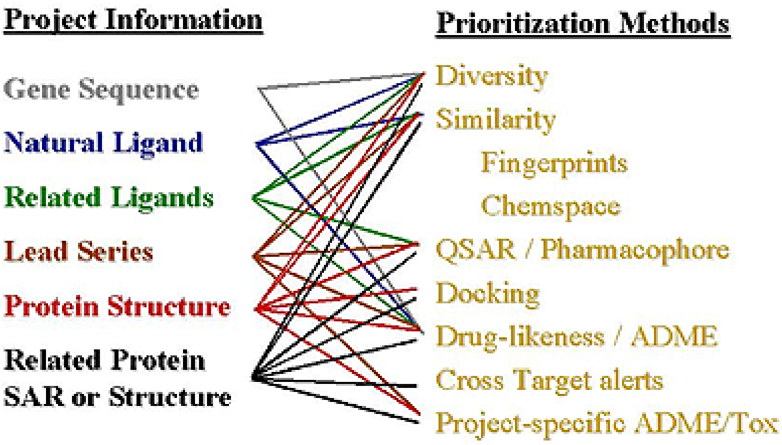
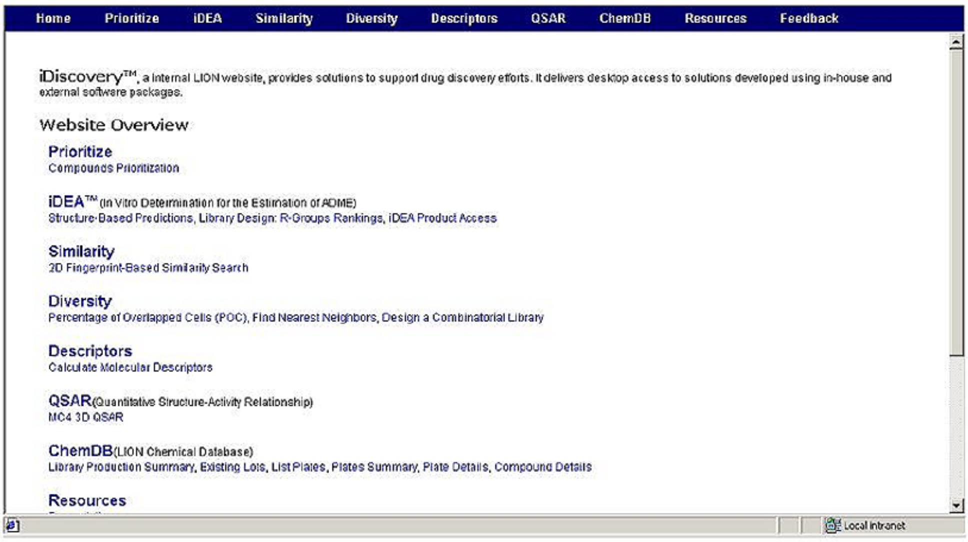


Figure 4. An example of an "available data / available application" matrix

ACCESSING INTEGRATED DATA

Once an integrated data/query system has been created, it must be delivered to the end user in such a way that it actually enhances their work. Here the major challenge is a "people issue". The interface must be simple and familiar, and should not require too much specific additional training for the user to start using it. In our solution to this problem, we have chosen web pages as an interface since it is familiar. has simple interfaces, and is easy to access. Figure 5 represents a "front door" to the system, the place where the end user chooses what task they want to accomplish. Figure 6 demonstrates a simple interface for searching multiple databases simultaneously using an sd, molfile, or sketched molecule.



iDiscovery™, a internal LION website, provides solutions to support drug discovery efforts. It delivers desktop access to solutions developed using in-house and external software packages.

Website Overview

Prioritize
Compounds Prioritization

iDEA™ (In Vitro Determination for the Estimation of ADME)
Structure-Based Predictions, Library Design: R-Groups Rankings, iDEA Product Access

Similarity
2D Fingerprint-Based Similarity Search

Diversity
Percentage of Overlapped Cells (POC), Find Nearest Neighbors, Design a Combinatorial Library

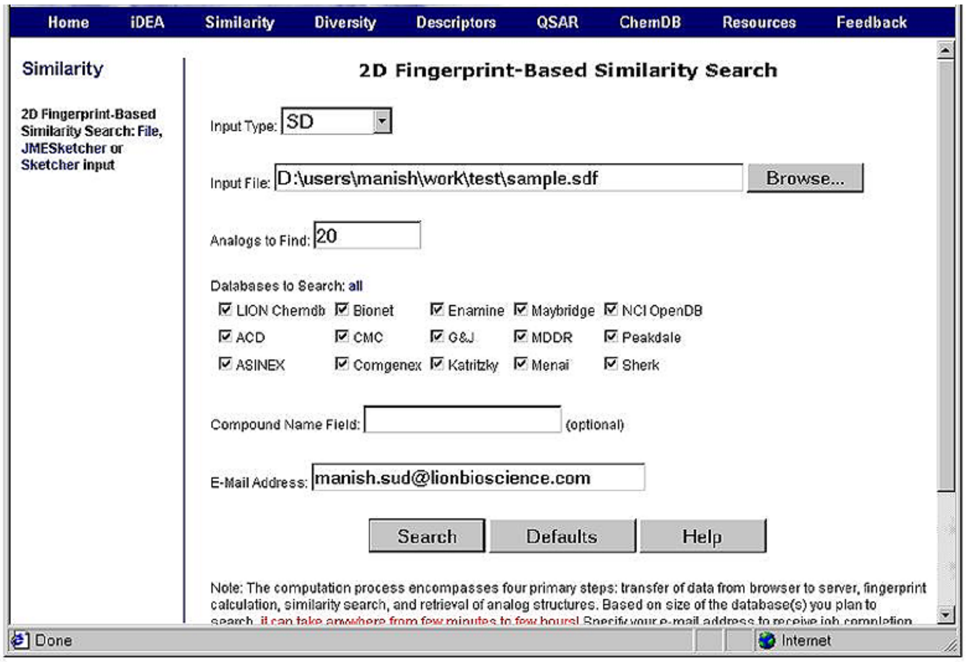
Descriptors
Calculate Molecular Descriptors

QSAR(Quantitative Structure-Activity Relationship)
MC4 3D QSAR

ChemDB(LION Chemical Database)
Library Production Summary, Existing Lots, List Plates, Plates Summary, Plate Details, Compound Details

Resources

Figure 5. Simple "front page" access to a set of integrated data, models, and tools.



Similarity

2D Fingerprint-Based Similarity Search: File, JMESketcher or Sketcher input

2D Fingerprint-Based Similarity Search

Input Type: SD

Input File: D:\users\manish\work\test\sample.sdf

Analogs to Find: 20

Databases to Search: all

☑ LION Chemdb	☑ Bionet	☑ Enamine	☑ Maybridge	☑ NCI OpenDB
☑ ACD	☑ CMC	☑ G&J	☑ MDDR	☑ Peakdale
☑ ASINEX	☑ Comgenex	☑ Katritzky	☑ Menai	☑ Sherk

Compound Name Field: (optional)

E-Mail Address: manish.sud@lionbioscience.com

Note: The computation process encompasses four primary steps: transfer of data from browser to server, fingerprint calculation, similarity search, and retrieval of analog structures. Based on size of the database(s) you plan to search, it can take anywhere from few minutes to few hours! Specify your e-mail address to receive job completion

Figure 6. Simple interface for performing a 2D similarity search of multiple databases simultaneously.

Turning Data into Sharable Knowledge Using Computational Tools

Figure 4 shows the computational methods available given a starting set of experimental data. These approaches have been powerful in enabling complicated hypothesis-driven experimental design in drug discovery project. However, the resultant models are usually "put on the shelf' as projects are promoted or discontinued, leaving these synthesised data unused for future projects. Having these computational models available, and using them appropriately could prove very valuable in addressing specificity, ADME, and safety information of current and future projects, especially in the case of organisations pursuing the target-class approach to drug discovery. Figure 7 shows a matrix of potency/specificity receptor-relevant chemspace (Pearlman reference) models created for a collection of nuclear receptors, representing a collection of easily applicable knowledge about a large fraction of proteins in the nuclear receptor target family. Though we have presented receptor-relevant chemspace models as our in silico surrogates for potency/selectivity, any computationally derived model, from similarity clustering methods to docking may be assembled, integrated, and applied in a cross-project manner.

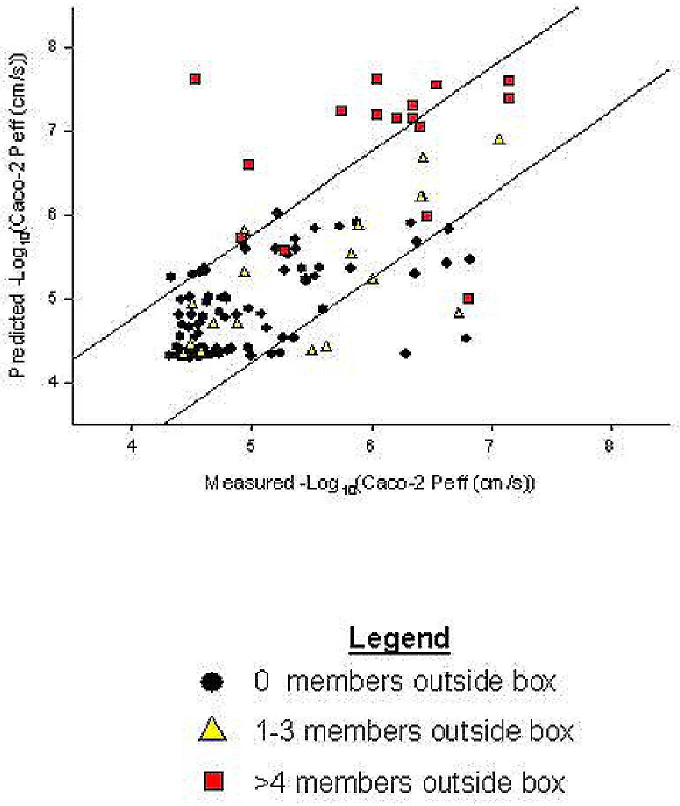
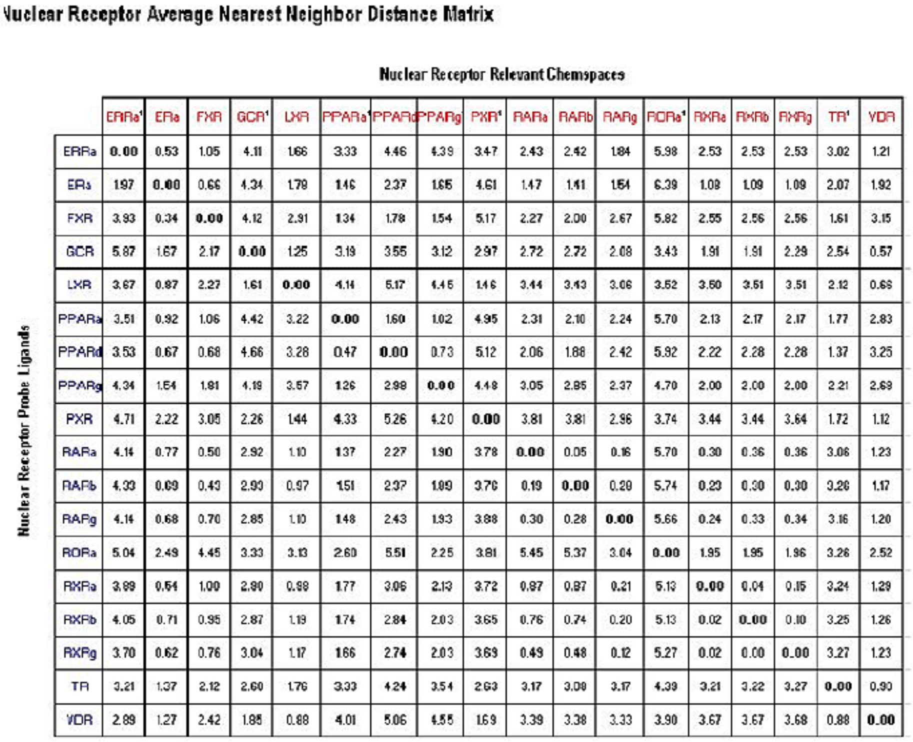


Figure 7. Distance matrix representing the relatedness between receptor relevant chemspace models.

Appropriately Using Data in the Form of Computational Models

Having data available in an integrated, searchable, analysable context, should be very valuable, however, more value could be added to this data by the creation of robust computational summaries (models) of the data, and applying them in appropriate ways. As an example, may varied approaches and algorithms exist that take a set of compounds and experimental activity data and derive a QSAR model that can often accurately predict the potential activity of a new compound.

It should be possible to reuse these models as a component of a knowledge environment for in silico evaluation of every compound against a surrogate for every potential experimental assay. However, most of the approaches used to build these computational models are statistical in nature, and therefore the performance of these models is only interpolative in nature, therefore, the models will likely perform poorly outside of "chemical space" (extrapolation) they were built upon. An approach to addressing this problem is illustrated in Figure 8 which shows the results from a model to predict Caco-2 effective permeability, including measures of uncertainty, using only the chemical structure of the compound as the only input. This caco-2 model was built using sophisticated statistical pattern recognition methods in which the output is a consensus prediction from 10 independent models. Each model is trained on a different representation of chemical space. To calculate the prospective measure of uncertainty (M.O.U) for the model, the upper and lower bounds of the chemical features (chemical descriptors) were determined to represent the bounds of the multidimensional chemical space upon which each model was trained. For each of the 10 independent (child) models, a new compound may have features whose values are outside of he bounds of the training set.



Nuclear Receptor Average Nearest Neighbor Distance Matrix

Nuclear Receptor Relevant Chemspaces

Nuclear Receptor Probe Ligands

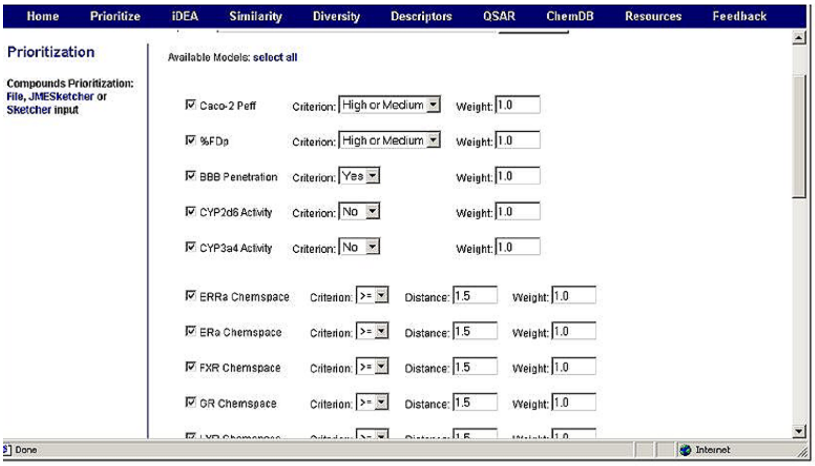
	ERRa'	ERa	FXR	GCR'	LXR	PPARa'	PPARd	PPARg	PXR'	RARa	RARb	RARg	RORa'	RXRa	RXRb	RXRg	TR'	VDR
ERRa	**0.00**	0.53	1.05	4.11	1.66	3.33	4.46	4.39	3.47	2.43	2.42	1.84	5.98	2.53	2.53	2.53	3.02	1.21
ERa	1.97	**0.00**	0.66	4.34	1.79	1.46	2.37	1.65	4.61	1.47	1.41	1.54	6.39	1.09	1.09	1.09	2.07	1.92
FXR	3.93	0.34	**0.00**	4.12	2.91	1.34	1.78	1.54	5.17	2.27	2.00	2.67	5.82	2.55	2.56	2.56	1.61	3.15
GCR	5.87	1.67	2.17	**0.00**	1.25	3.19	3.55	3.12	2.97	2.72	2.72	2.08	3.43	1.91	1.91	2.29	2.54	0.57
LXR	3.67	0.87	2.27	1.61	**0.00**	4.14	5.17	4.45	1.46	3.44	3.43	3.06	3.52	3.50	3.51	3.51	2.12	0.66
PPARa	3.51	0.92	1.06	4.42	3.22	**0.00**	1.60	1.02	4.95	2.31	2.10	2.24	5.70	2.13	2.17	2.17	1.77	2.83
PPARd	3.53	0.67	0.68	4.66	3.28	0.47	**0.00**	0.73	5.12	2.06	1.88	2.42	5.92	2.22	2.28	2.28	1.37	3.25
PPARg	4.34	1.64	1.81	4.19	3.57	1.26	2.98	**0.00**	4.48	3.05	2.85	2.37	4.70	2.00	2.00	2.00	2.21	2.69
PXR	4.71	2.22	3.05	2.26	1.44	4.33	5.26	4.20	**0.00**	3.81	3.81	2.96	3.74	3.44	3.44	3.64	1.72	1.12
RARa	4.14	0.77	0.50	2.92	1.10	1.37	2.27	1.90	3.78	**0.00**	0.05	0.16	5.70	0.30	0.36	0.36	3.06	1.23
RARb	4.33	0.69	0.43	2.93	0.97	1.51	2.37	1.89	3.76	0.19	**0.00**	0.29	5.74	0.23	0.30	0.30	3.26	1.17
RARg	4.14	0.68	0.70	2.85	1.10	1.48	2.43	1.93	3.88	0.30	0.28	**0.00**	5.66	0.24	0.33	0.34	3.16	1.20
RORa	5.04	2.49	4.45	3.33	3.13	2.60	5.51	2.25	3.81	5.45	5.37	3.04	**0.00**	1.95	1.95	1.96	3.26	2.52
RXRa	3.89	0.64	1.00	2.90	0.98	1.77	3.06	2.13	3.72	0.87	0.87	0.21	5.13	**0.00**	0.04	0.15	3.24	1.29
RXRb	4.05	0.71	0.95	2.87	1.19	1.74	2.84	2.03	3.65	0.76	0.74	0.20	5.13	0.02	**0.00**	0.10	3.25	1.26
RXRg	3.70	0.62	0.76	3.04	1.17	1.66	2.74	2.03	3.69	0.49	0.48	0.12	5.27	0.02	0.00	**0.00**	3.27	1.23
TR	3.21	1.37	2.12	2.60	1.76	3.33	4.24	3.54	2.63	3.17	3.09	3.17	4.39	3.21	3.22	3.27	**0.00**	0.90
VDR	2.89	1.27	2.42	1.85	0.88	4.01	5.06	4.55	1.69	3.39	3.38	3.33	3.90	3.67	3.67	3.68	0.88	**0.00**

Figure 8. Example application of prospective measures of uncertainty in a predictive model for caco-2 effective permeability.

If the features are outside of the bounds for any feature for a child model, that model is considered to be extrapolating. Figure 7 shows an example of the consequences of using extrapolated increases, the error in prediction increases dramatically. Developing methods to assess prediction confidence for all models could aid dramatically both in their successful first-time use as well as enable appropriate reuse of the knowledge gleaned by their creation.

Simultaneous Use of Computational Models and Integrated Data for Compound Prioritisation and Assessment

During Lead Identification and Lead Optimisation, prioritisation or ranking of compounds to acquire, plate, or synthesise can be cumbersome, and is often performed without using all available information. Similarly, assessment of which compound or compound series to promote to the next phase of research should also be performed using all available information simultaneously. Figure 9 shows a simple compound prioritisation input screen for simultaneous prediction of ADME and potency/specificity properties. In this input screen the user may choose which properties to calculate, the criteria defining whether a given compound passes or fails, and the weight of that property in the calculation of a summary score for a molecule.



Prioritization

Compounds Prioritization: File, JMESketcher or Sketcher input

Available Models: select all

☑ Caco-2 Peff Criterion: High or Medium Weight: 1.0

☑ %FDp Criterion: High or Medium Weight: 1.0

☑ BBB Penetration Criterion: Yes Weight: 1.0

☑ CYP2d6 Activity Criterion: No Weight: 1.0

☑ CYP3a4 Activity Criterion: No Weight: 1.0

☑ ERRa Chemspace Criterion: >= Distance: 1.5 Weight: 1.0

☑ ERa Chemspace Criterion: >= Distance: 1.5 Weight: 1.0

☑ FXR Chemspace Criterion: >= Distance: 1.5 Weight: 1.0

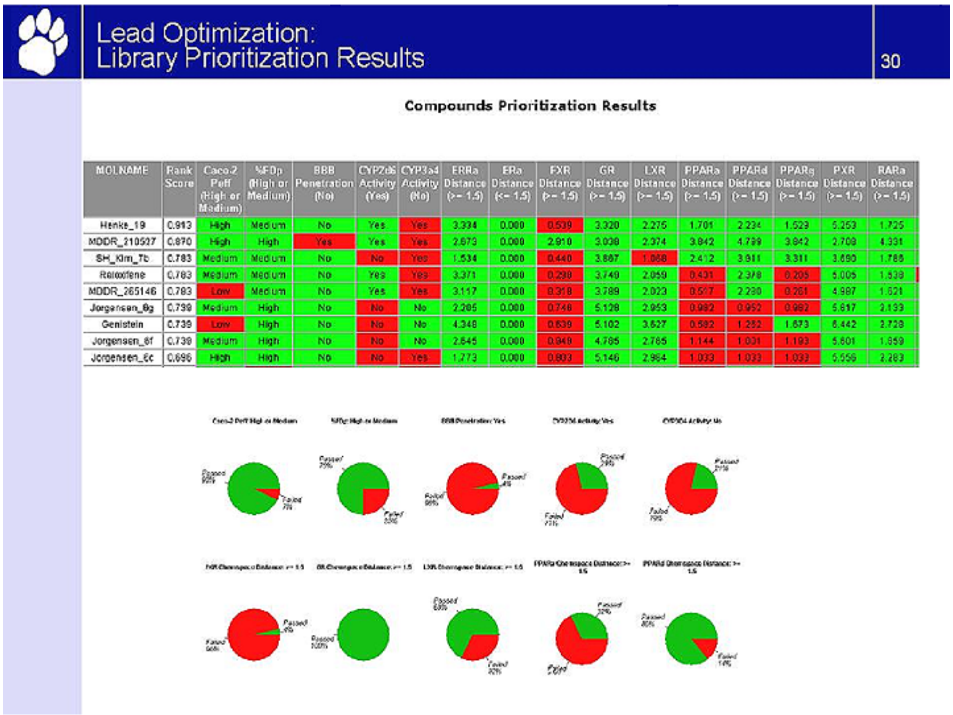
☑ GR Chemspace Criterion: >= Distance: 1.5 Weight: 1.0

Figure 9. Prioritisation Input Screen. The end user selects the models to run, the criteria required to pass, and the weight each model contributes to the final score.

The summary score allows the composite ranking of all compound under consideration using the same objective evaluation criterion. Figure 10 shows the results of a prioritisation in two views. First, a compound-by-compound view, and second, a property distribution view, which would likely be useful for selecting from commercially available compound collections of in the evaluation of whether or not to synthesise a particular series of compounds versus another. By including integrated data determined experimentally in this analysis, compoung assessment can also be performed.

Application of Integrated Approaches as a Cheminformatics Solution for Decision Support in Drug Discovery: A Hypothetical Estrogen Receptor Project

For this hypothetical example, we will assume that we have chosen Estrogen Receptor alpha (Era) as the target of our Lead Identification (LI)/Lead Optimisation (LO) project. We perform a search of the literature, and find several examples of compounds which have been shown to interact with Era (Figure 11).



Lead Optimization:
Library Prioritization Results

30

Compounds Prioritization Results

MOLNAME	Rank Score	Caco2 Peff (High or Medium)	%FDp (High or Medium)	BBB Penetration (No)	CYP2d6 Activity (Yes)	CYP3a4 Activity (No)	ERRa Distance (>= 1.5)	ERa Distance (<= 1.5)	FXR Distance (>= 1.5)	GR Distance (>= 1.5)	LXR Distance (>= 1.5)	PPARa Distance (>= 1.5)	PPARd Distance (>= 1.5)	PPARg Distance (>= 1.5)	PXR Distance (>= 1.5)	RARa Distance (>= 1.5)
Henke_19	0.913	High	Medium	No	Yes	Yes	3.334	0.000	0.539	3.320	2.275	1.701	2.234	1.529	5.253	1.725
MDDR_210527	0.870	High	High	Yes	Yes	Yes	2.673	0.000	2.910	3.038	2.374	3.842	4.799	3.842	2.708	4.331
SH_Kim_7b	0.783	Medium	Medium	No	No	Yes	1.534	0.000	0.440	3.887	1.068	2.412	3.911	3.311	3.890	1.788
Raloxifene	0.783	Medium	Medium	No	Yes	Yes	3.371	0.000	0.298	3.749	2.059	0.431	2.378	0.205	5.005	1.538
MDDR_265146	0.783	Low	Medium	No	Yes	Yes	3.117	0.000	0.318	3.789	2.023	0.517	2.280	0.261	4.987	1.621
Jorgensen_6g	0.739	Medium	High	No	No	No	2.205	0.000	0.748	5.128	2.953	0.982	0.952	0.982	5.617	2.133
Genistein	0.739	Low	High	No	No	No	4.348	0.000	0.639	5.102	3.627	0.682	1.252	1.673	6.442	2.728
Jorgensen_6f	0.739	Medium	High	No	No	No	2.645	0.000	0.948	4.785	2.785	1.144	1.031	1.193	5.601	1.859
Jorgensen_6c	0.696	High	High	No	No	Yes	1.773	0.000	0.803	5.146	2.984	1.033	1.033	1.033	5.556	2.283

Caco-2 Peff High or Medium
%Fdp: High or Medium
BBB Penetration: Yes
CYP2D6 Activity: Yes
CYP3A4 Activity: No
FXR Chemspace Distance >= 1.5
GR Chemspace Distance >= 1.5
LXR Chemspace Distance >= 1.5
PPARa Chemspace Distance >= 1.5
PPARd Chemspace Distance >= 1.5
Passed
Failed

Figure 10. Output from compound prioritisation. Compound dependent rollup of predictions, and compound collection distributions of predictions.

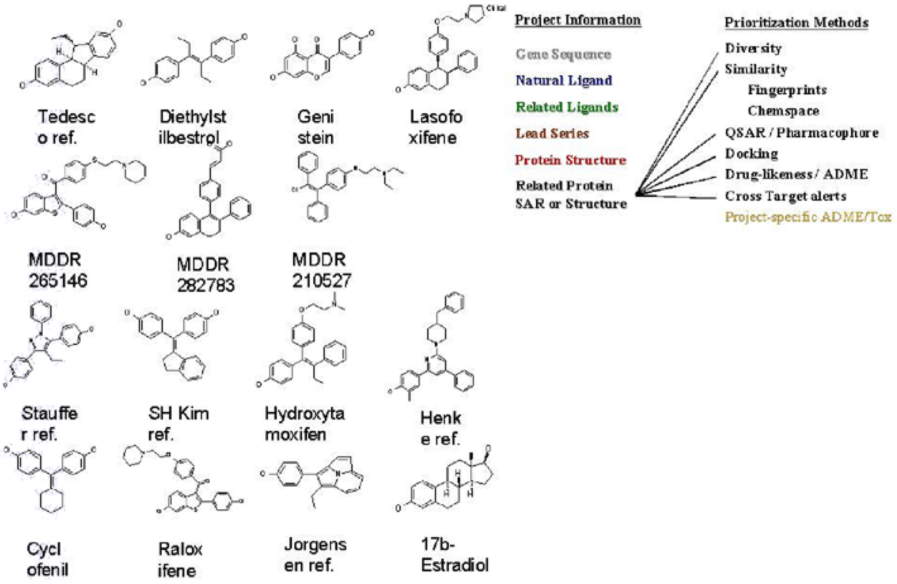
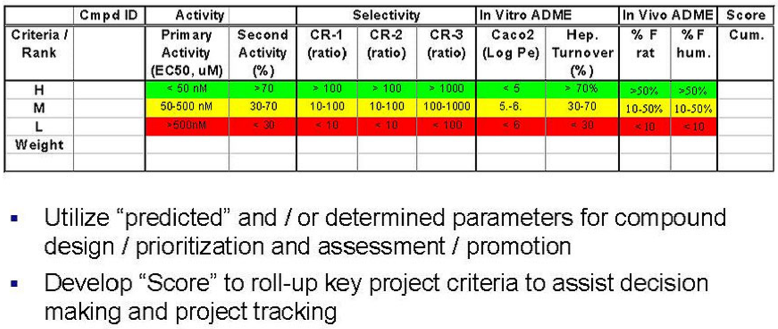


Figure 11. Starting information for the hypothetical ER-a project.

In addition, since our organisation is using a target class approach to drug discovery, we also retrieve all known NR-ligand pairs from the literature as an initial knowledge environment. ER-a is an example of a data-rich project (Figure 11), and therefore almost every method available for computational model generation would be at our disposal. Once the project criteria and starting assumptions (Example shown in Figure 12) have been established, lead identification

(Figure 3) can begin. A starting potential screening set is selected by a search of the known ligands against all available compounds (both virtual and existing - either in house or available from vendors - example shown in Figure 6).

This screening set is further reduced by simultaneous parameter evaluation using integrated computational tools. The project team decides where to cut off the screening collection, and these compounds are assembled, synthesised or acquired, and are run in initial potency assays.



	Cmpd ID	Activity		Selectivity			In Vitro ADME		In Vivo ADME		Score
Criteria / Rank		Primary Activity (EC50, uM)	Second Activity (%)	CR-1 (ratio)	CR-2 (ratio)	CR-3 (ratio)	Caco2 (Log Pe)	Hep. Turnover (%)	% F rat	% F hum.	Cum.
H		< 50 nM	>70	> 100	> 100	> 1000	< 5	> 70%	>50%	>50%	
M		50-500 nM	30-70	10-100	10-100	100-1000	5.-6.	30-70	10-50%	10-50%	
L		>500nM	< 30	< 10	< 10	< 100	< 6	< 30	<10	<10	
Weight											

- Utilize "predicted" and / or determined parameters for compound design / prioritization and assessment / promotion
- Develop "Score" to roll-up key project criteria to assist decision making and project tracking

Figure 12. Project criteria and use example.

In this scenario, all evaluations can be performed by any and all members of the project team through a web site with common default settings for project criterion. After the screening results are returned and confirmed, the project team then must assess where the project is at that point in time, and decide where to focus the available resources. In this example, three series of compounds passed the requirements necessary for lead optimisation (Figure 13), however, the project team has only enough synthetic resources to work on one series at at time, therefore, the project team must decide which series has the most potential for success.

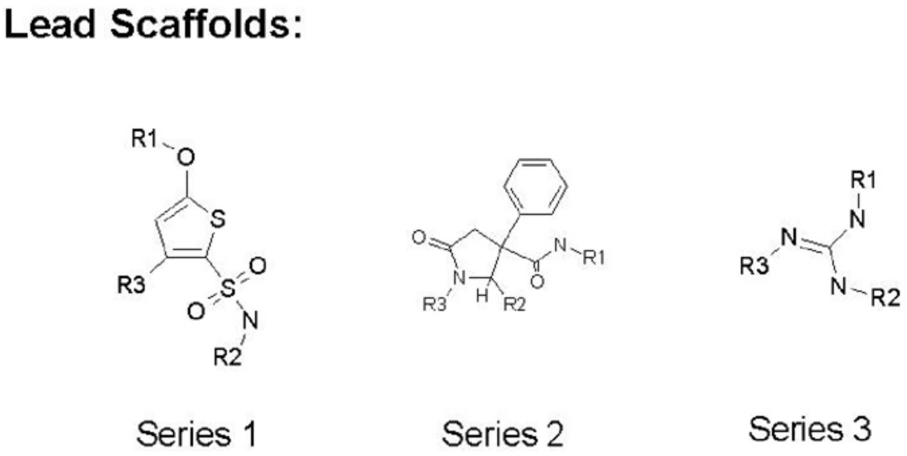


Figure 13. Hypothetical leads to choose between for prioritisation for optimisation.

To evaluate the potential of each series, virtual libraries will be enumerated, and the resultant products will be evaluated simultaneously (Figure 9). Then, to rate the potential of the libraries in the context of the other libraries, the distributions the components are compared simultaneously between the libraries representing the three scaffolds. Shown in Figure 14, this analysis can be used as an evaluation of potential liabilities which are best compared by analysing the fraction of each library that fails the project criteria for that component.

In this analysis, series 1 has the fewest bad distribution of liabilities, and should be chosen by the project team for further research, with series 2 representing a potential backup series.

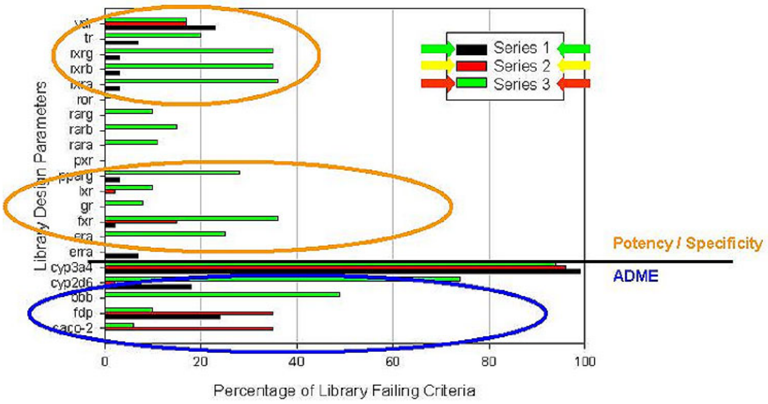
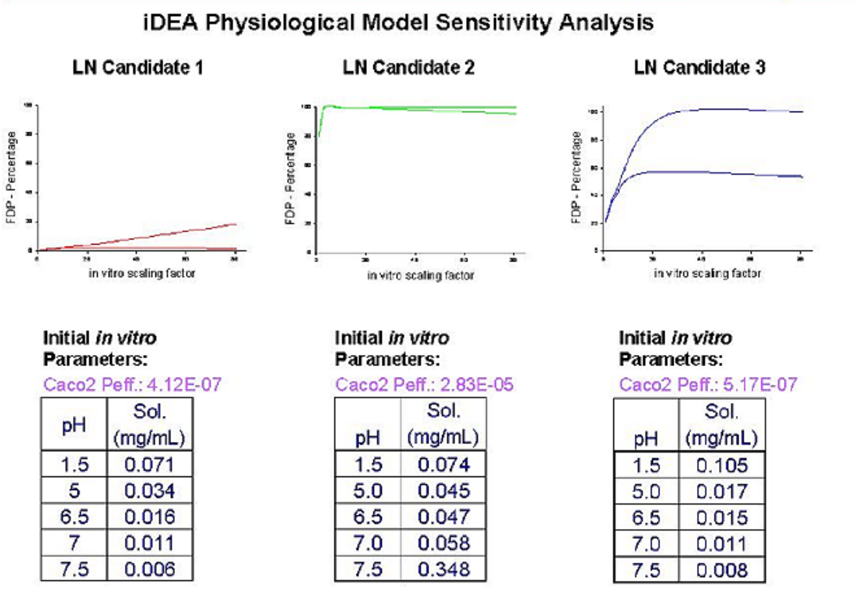


Figure 14. Simultaneous library property distribution comparison.

As the project moves closer to lead selection fpr pre-clinical development, the data integration components start to take priority over the in silico predictions. However, computational approches still have tremendous value at this stage, allowing the project team to evaluate the potential performance of candidate molecules in man. Figure 15 shows an analysis of the human absorption potential for three compounds representative of candidate series, in which the experimentally determined solubility and permeability values have been varied systematically.

As can be seen in the graphs, three classes of behaviors can be observed. In one example neither increases in solubility nor permeability can increase the absorption potential, which remains relatively low. In the second example, increases in solubility or permeabiltiy also have no effect on the human absorption potential, which is high. In the final example, increases in solubility and permeability show marked changes in the human absorption potential. The candidate molecules from the second and third examples are therefore the compound series to bring forward if all other factors besides absorption are equal. In addition, this analysis illustrates which series deserve followup study for the development of second generation compounds.



Initial *in vitro* Parameters:
Caco2 Peff.: 4.12E-07

pH	Sol. (mg/mL)
1.5	0.071
5	0.034
6.5	0.016
7	0.011
7.5	0.006

Initial *in vitro* Parameters:
Caco2 Peff.: 2.83E-05

pH	Sol. (mg/mL)
1.5	0.074
5.0	0.045
6.5	0.047
7.0	0.058
7.5	0.348

Initial *in vitro* Parameters:
Caco2 Peff.: 5.17E-07

pH	Sol. (mg/mL)
1.5	0.105
5.0	0.017
6.5	0.015
7.0	0.011
7.5	0.008

Figure 15. Use of sensitivity analysis for the evaluation of human absorption potential for lead selection.

SUMMARY/CONCLUSION

So what is different about this approach? The drug discovery process is a fairly evolved one. Data has been shared and models developed for use in drug discovery process ever since affordable computers showed up in the marketplace.

However, the primary storage locale for experimental data is still Microsoft Excel, and the primary conduit of information is through individual interactions between project team members. Unfortunately, humans do not work with complete fidelity in serving as data or knowledge-sharing nodes. One could argue that modern drug discovery is using the ancient means of folklore as the method of knowledge sharing - this clearly is not sufficient given the ever exploding number of targets, hits, and interactions. This paper has described the initial steps of building a system that uses integrated databases to store *all* the data determined for discovery projects. It also indicates that this data is best used in synthesised from through the appropriate application of computational model building and their resultant use. Finally it illustrates the potentional power of combining the data and computational models in a system that allows the end user to simultaneously consider all available properties, and therefore make more and presumable better decisions about prioritising resources in the support of drug discovery and development.