Chemistry & Biology, Vol. 11, 1495–1503, November, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.chembiol.2004.08.026

A Flexible Data Analysis Tool for Chemical Genetic Screens

Brian P. Kelley,[1] Mitchell R. Lunn,[1] David E. Root,[3] Stephen P. Flaherty,[1] Allison M. Martino,[1] and Brent R. Stockwell[1,2,*]
[1]Department of Biological Sciences
[2]Department of Chemistry
Columbia University
Fairchild Center, MC 2406
1212 Amsterdam Avenue
New York, New York 10027

Summary

High-throughput assays generate immense quantities of data that require sophisticated data analysis tools. We have created a freely available software tool, SLIMS (*S*mall *L*aboratory *I*nformation *M*anagement *S*ystem), for chemical genetics which facilitates the collection and analysis of large-scale chemical screening data. Compound structures, physical locations, and raw data can be loaded into SLIMS. Raw data from high-throughput assays are normalized using flexible analysis protocols, and systematic spatial errors are automatically identified and corrected. Various computational analyses are performed on tested compounds, and dilution-series data are processed using standard or user-defined algorithms. Finally, published literature associated with active compounds is automatically retrieved from Medline and processed to yield potential mechanisms of actions. SLIMS provides a framework for analyzing high-throughput assay data both as a laboratory information management system and as a platform for experimental analysis.

Introduction

Chemical genetics is an emerging approach for studying biological processes [1–8]. In this genetic-like screening approach, thousands of small organic molecules are tested for activity in protein-targeted, cellular, or organismal assays. Subsequent studies use active compounds to link phenotypic changes in cells or organisms to the modulation of specific proteins' functions. Thus, in this approach, organic compounds are used as tools for determining the macromolecules that regulate cellular and organismal phenotypes.

Chemical genetics optimally involves active collaborations between chemists and biologists. This is in contrast to some large drug discovery organizations, in which high-throughput screening operations exist as independent service facilities. Interactions between chemists, biologists, and automation specialists in such organizations may be limited to the nomination of an assay by biologists, screening by automation specialists, and transfer of compounds to chemists for subsequent optimization. Successful chemical genetic projects, in contrast, typically involve active and continuous sharing of data between chemists, biologists, and automation specialists, collaborative refinement of screening and analysis protocols, and ongoing testing and retesting of compounds. Thus, a flexible data analysis tool that allows data sharing is essential for collaborative chemical genetic research.

Moreover, the number of compounds tested in chemical screens ranges from several thousand up to a million. In order to extract the maximal amount of information from such screens, it is necessary to store electronic chemical structures for all the tested compounds and to create an automated method of processing the resulting assay data. These data are typically obtained in the form of a large number of raw plate-reader files indicating the level of fluorescence, absorption, or luminescence in each microtiter plate well.

Thus, a critical issue for investigators performing chemical genetic screens is information management and data sharing [9–25]. While a number of commercial systems have been developed to organize and to analyze large volumes of screening data for industrial organizations, such systems have fatal limitations for collaborative chemical genetic investigations, especially those in academia [26–32]. First, existing systems are not designed to enable collaboration between different institutions and laboratories or to deliver searchable content for the purpose of publication. Second, they are geared toward static experimental design, and do not allow biologists and chemists to rapidly modify experimental and analytic procedures. Third, they do not allow incorporation of novel analytic modules as they are developed, such as automated detection of systematic errors or automatic querying of databases to reveal potential mechanisms of action of compounds. The net effect of these deficiencies is that collaborative projects that involve a major screening component do not have a readily available software option for managing and analyzing high-throughput chemical screening data.

Moreover, information management systems are needed for controlling aspects of high-throughput experimental design and analysis [14, 15, 25–27, 29, 30]. When performing high-throughput experiments that are susceptible to artifacts induced by minor experimental variations, it is essential to annotate and to document precisely experimental methods (e.g., equipment usage and lot numbers of reagents [33–37]).

Finally, artifacts that are generated through systematic error need to be corrected to ensure proper reporting of biological results.

In this report, we describe the use of SLIMS in a chemical genetic screen related to spinal muscular atrophy (SMA). SMA is an autosomal recessive disease involving degeneration of α-motor neurons in the spinal cord anterior horn, leading to progressive muscular atrophy, paralysis, respiratory failure, and infant death. SMA

*Correspondence: stockwell@biology.columbia.edu
[3]Present address: Broad Institute of Harvard and Massachusetts Institute of Technology, 320 Charles Street, Cambridge, Massachusetts 02141.

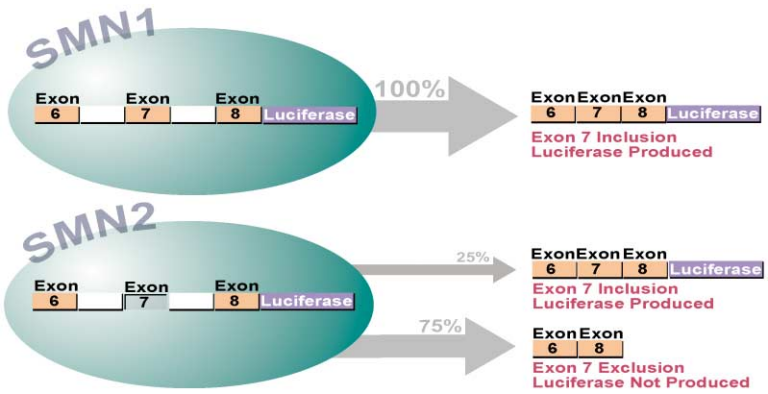

Figure 1. Molecular Genetics of Spinal Muscular Atrophy

Spinal muscular atrophy (SMA) is an autosomal recessive disease involving degeneration of α-motor neurons in the spinal cord anterior horn, leading to progressive muscular atrophy, paralysis, respiratory failure, and infant death. SMA is caused by deletion of the survival motor neuron 1 (*SMN1*) gene, which encodes the SMN protein.

Humans have two copies of the *SMN* gene (*SMN1* and *SMN2*), which are located in a 500 kilobase (kb) inverted repeat on chromosome 5q13. In SMA patients, the *SMN1* gene is deleted entirely and *SMN2* is spliced such that only ~25% of mRNAs are full length, and ~75% of spliced transcripts exclude exon 7. The *SMN2* gene product is sufficient to maintain fetal development, but affected individuals manifest the symptoms of SMA early in life and typically die as infants.

is caused by deletion of the survival motor neuron 1 (*SMN1*) gene, which encodes the SMN protein. Humans have a related copy of the *SMN* gene, denoted *SMN2*. Both are located in a 500 kilobase (kb) inverted repeat on chromosome 5q13. In SMA patients, the *SMN1* gene is deleted entirely and *SMN2* is spliced such that only ~25% of mRNAs are full length, while ~75% of spliced transcripts exclude the required exon 7. The product of the *SMN2* gene is sufficient to maintain fetal development, but affected individuals manifest the symptoms of SMA early in life and typically die as infants.

As it has been previously reported that the lack of exon 7 is the major cause of the *SMN2* gene producing a nonfunctional, truncated SMN protein (Figure 1) [38], we developed a phenotype-based assay designed to detect proper mRNA splicing. In our study, we used an immortalized human cervical carcinoma cell line stably expressing a *SMN*-minigene-reporter reporter system [39] that consists of exons 6 through 8 with intervening introns of either *SMN1* or *SMN2*, which we obtained from Androphy and Zhou (University of Massachusetts Medical School, Worcester, MA). This construct also contained a gene conferring neomycin resistance, which guaranteed that all cells growing in medium containing G418 possessed the minigene construct. Luciferase, the reporter gene's product, is only produced when proper splicing (i.e., incorporation of exon 7 in the final mRNA transcript) occurs. Exclusion of exon 7 results in the luciferase gene being out-of-frame and is therefore not translated.

Over 47,000 compounds were screened using this assay: 20,000 compounds were from a combinatorial chemistry library from ComGenex; 1,040 compounds were from a National Institute for Neurological Disorders and Stroke library; 2,337 compounds were from our Annotated Compound Library [40]; and 23,685 were from our TIC Library. The TIC Library is a composite of compounds available from TimTec, IBS, and ChemBridge that were selected for specific properties, including natural-product likeness.

Due to the format of our assay, we were required to study compounds that produced an effect that was selective for the *SMN2*-minigene-reporter cells, as non-selectivity suggests that a compound could be affecting the promoter of the construct (which is not the same *SMN* promoter found in human cells) or that it could be modulating the growth rate of the cells. To this end, our compound libraries were also tested on *SMN1*-minigene-reporter cells. We refer to these cell types as *SMN2*-LUC and *SMN1*-LUC.

Results

We loaded electronic structures for 47,062 compounds physically present and plated in our laboratory into SLIMS and recorded for each compound the vendor, initial amount purchased, and the location on each plate of each compound. We designed SLIMS to support the industry-standard structure data (SD) format. To load compound data, we performed an initial scan of each SD file to validate that the compounds were correctly labeled and contained the appropriate information. SLIMS can be configured to require mandatory data for each compound, such as the compound's vendor and catalog number.

After we loaded compounds and plate locations into the SLIMS database, we reformatted these electronic plates from 96-well plates (provided by the original vendors) to 384-well plates, which we used for screening operations. This electronic reformatting operation was accomplished through the use of a simple SLIMS wizard. Compound transfers from mother plates, which are delivered from the vendor, to daughter plates, which contain the stock compounds diluted in DMEM, and assay plates, which contain the compounds, and the cells diluted in DMEM are also performed in this fashion.

The SMA study was performed over several months and required considerable protocol optimization. To develop the assay, protocols for data analysis needed to be dynamic and flexible, as new analyses and experimental conditions were developed. SLIMS provides rapid methods for generating protocols and allows multiple protocols to be used while loading an assay into the database. This is especially important during the protocol optimization process, where old data should coexist with newer data from the same assay. SLIMS provides a

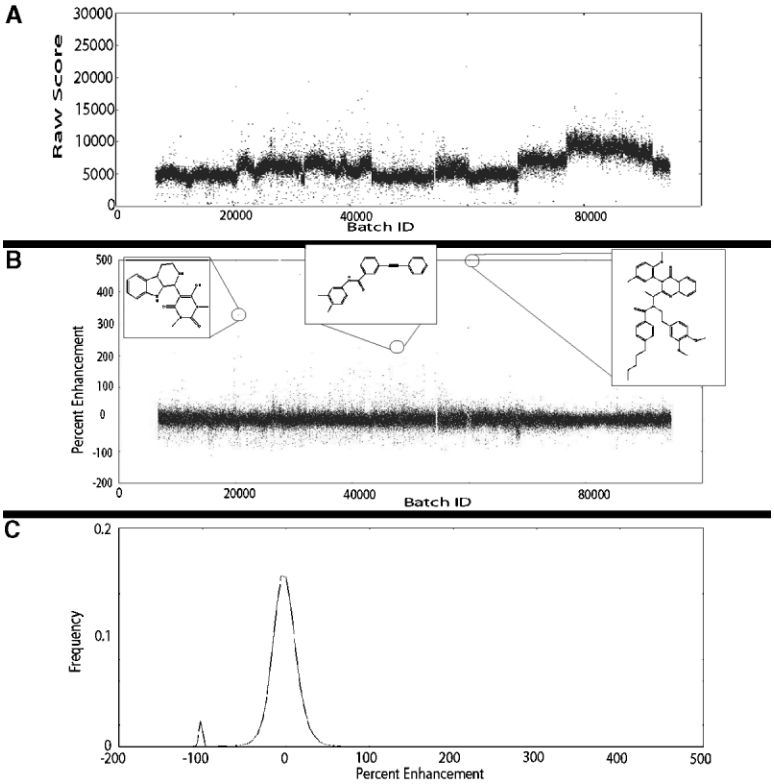

Figure 2. Normalization of Raw Data in SLIMS

(A) The SMA assay raw data. Notice the systematic drift in the raw data.

(B) The raw data are scored using the appropriate protocol descriptions. These data only show small drift. Each data point is linked to the compound so users can click on data and see the related compound.

(C) The histogram of the percent enhancement shows a standard assay response. The small bump on the left is the negative controls.

protocol-creation wizard to indicate the location of positive and negative controls and to select the scoring scheme for normalizing each plate. Scoring systems built into SLIMS include standard "percent enhancement" and "percent inhibition" computations. SLIMS is sufficiently flexible that experiments can be loaded with multiple protocols. This ensures that, if the protocol changes during an experiment, all the experimental data can be analyzed as a whole. As newly created assay data is added to SLIMS, they can be immediately visualized in order to facilitate the optimization process. This is especially important for detecting and rejecting failed plates.

Visualization of the data generated over time shows a significant drift in raw luminescence values for the SMA screen (Figure 2). This drift was caused by experimental conditions changing from day to day (e.g., light changes, evaporation). This drift emphasizes that placing control data on each plate is a necessity for normalizing data taken over long periods of time. Using these controls, the normalized data exhibits much less drift (Figure 2). Percent enhancement of each well compared to the untreated control wells that were placed on each plate (Figure 2) as follows:

$$\text{percent enhancement} = 100 \times \left(\frac{x - \text{negative}}{\text{positive} - \text{negative}} - 1 \right),$$

where *positive* are the untreated controls wells and *negative* are wells containing only DMSO. The plots generated by SLIMS are tied into the database, and clicking on a data point highlights the relevant compound and provides a quick way to scan for initial lead compounds.

Screening results, such as those obtained in this SMA screen, typically exhibit a continuous range of activity with a Gaussian distribution. There are several methods of selecting active compounds from such a large-scale screen. In the threshold approach, a cutoff value is chosen for the selection of hits, and the active compounds are confirmed in a repeat experiment, typically involving a dose-response curve. The cutoff criteria for determining hits may be based on absolute activity (i.e., 2-fold activity versus control), the distribution (i.e., three standard deviations or greater from the mean), or a desired number of compounds to be retested. Other methods use a cutoff to select the most active compounds and then use techniques such as locating structural analogs to add to the secondary screening. The SMA screening data showed a standard Gaussian distribution, allowing thresholds to be chosen for hit selection (Figure 2C).

Before the assay results could be analyzed, however, we needed to assess the quality of the screening data. Due to the nature of the luminescence-based screen, the screening data tended to exhibit spatial artifacts. These involved row effects, in which certain rows were dimmer than others, and edge effects, in which the edge of the plate tended to be brighter than the center. This can dramatically affect screening results due to the fact that plate controls are located at the edges. Edge effects and other systematic errors are caused by some repetitive fault in the instrumentation. Indeed, a closer look at the histogram in Figure 2C shows that the mean is less than zero, namely −3.8 percent enhancement. In our assay, the expected mean percent enhancement of all tested compounds should be zero net effect in that most tested drugs should not either enhance or inhibit

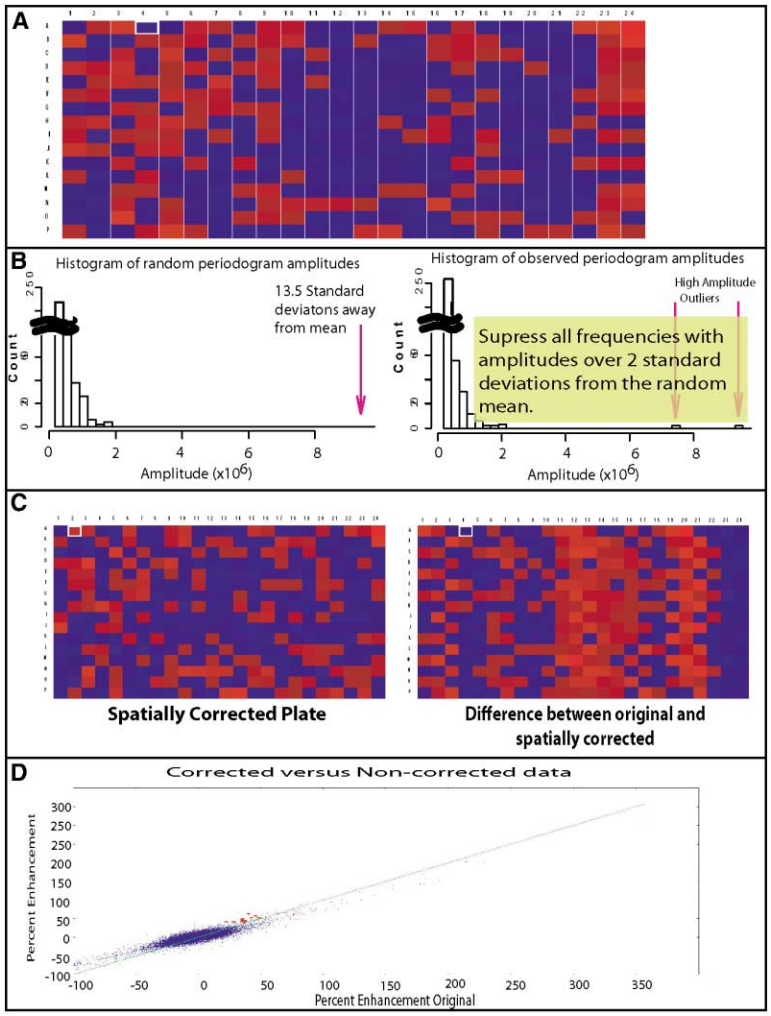

Figure 3. SLIMS Corrects for Spatial Error by Determining the Outlying Periodogram Frequency Amplitudes that Should Not Exist If the Data Displayed a Random Spatial Distribution

(A) The original plate with quite obvious plate effect. Red wells are the highest in the plate and blue values are the lowest.

(B) The distribution of peridogram amplitudes and comparison of the random distribution to the distribution detected in the plate. As indicated, all frequencies that appear as outliers are dampened to their values as if they had been generated by a random process.

(C) The spatially corrected plate on the left looks considerably better than the original. Furthermore, the difference plate is shown adjacent and indicates how the systematic error was corrected. Blue values have been suppressed in the original plate, and red values have been enhanced.

(D) The overall correction between the original and corrected data. The green line constitutes no change for the corrected data. Points marked in red have been restored as potential hits through the data-correction procedure.

the SMN-minigene reporters. This bias, even though subtle and hard to detect by eye, may occlude potential lead compounds.

We used a periodogram technique to automatically detect systematic errors [36]. Given a microtiter plate seeded with random compounds and no robotic errors, spatial patterns are unlikely and, indeed, unexpected. SLIMS computes a spatial randomness probability (P_{random}) for each assay plate. P_{random} is a qualitative measure of the amount of nonrandom spatial patterning appearing on the plate. For example, a P_{random} value of 0.5 indicates that the spatial pattern of the observed plate would be recapitulated in one out of every two plates randomly generated with the same mean and standard deviation as the observed plate. Plates that have a $P_{random} \leq 0.05$ are flagged for user inspection. In other words, plates that have a one in twenty chance of forming patterns that could be observed randomly are highlighted for inspection. The plate in Figure 3 has systematic errors that are difficult to detect by eye. This difficulty, combined with the fact that, for some people, it is difficult to analyze hundreds of plates manually, suggests that an automated method can be useful to correct systematic errors.

Individual inspection is both desirable and appropriate for small numbers of plates, but when the number of plates is large, this approach is unwieldy. SLIMS facilitates high-throughput error detection by automatically correcting for detected spatial patterns. To correct data, we locate systematic patterns and remove them by reducing the power of these high-amplitude outliers (Figures 3A–3C). Thus, high periodogram amplitudes, which are based on a random model, are suppressed. This process makes the assumption that each plate has been generated by a random process. Each well is assumed to be independent of all other wells in the plate. Correlations discovered between wells are either real assay responses, such as control wells, or some systematic error pattern, such as a row or column effect. High periodogram amplitudes imply correlated wells, and by suppressing these periodogram amplitudes, SLIMS can automatically remove spatial correlations (Figure 3B). Figure 3C indicates the correction performed wherein red regions of the plate have been corrected by lowering the detected value, and blue regions indicate wells that have been increased in value. Because correcting systematic error might in itself reduce real experiment response, SLIMS stores both the corrected and uncor-

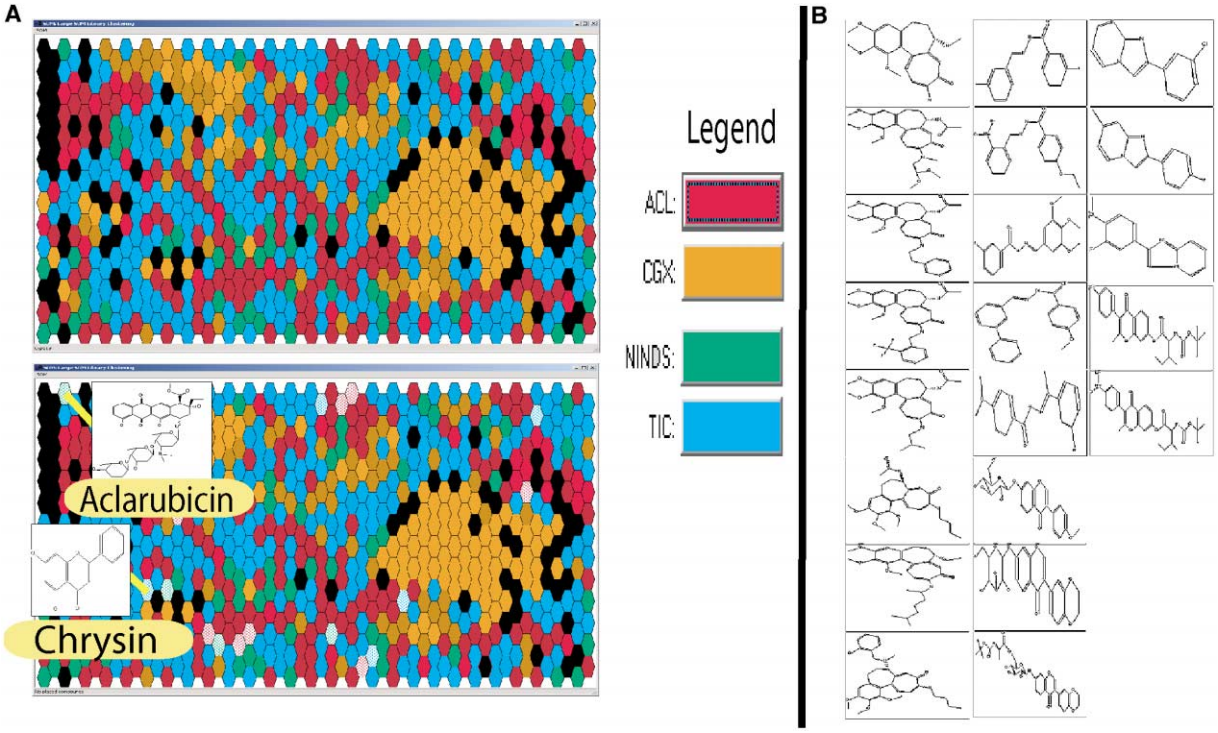

Figure 4. Diversity Analysis Using Self-Organizing Maps in SLIMS

(A) The self-organizing map generated through SLIMS using SLIMS database fingerprints. This SOM shows the structure-space location of *SMN2*-LUC SMA hits. Each node in the SOM contains a collection of similar compounds, and each neighboring node is more similar than nonneighboring nodes. Bright orange wells contain only compounds from the Comgenex (CGX) library, and dark orange wells contain mostly CGX compounds. The SMA hits are shown in the crosshatched wells in the lower map. Notice that the Comgenex library (colored orange) is clumped together in the SOM, which is indicative of a combinatorial library. The hits, however, are scattered around this representation of structural space and indicate that many different targets are potentially involved. Two representative active compounds, aclarubicin and chrysin, are marked on the lower map.

(B) Several classes of active analogs automatically generated from the SOM and by extracting scaffolds from lead compounds. These classes are used to both validate and prioritize lead compounds for dilution series.

rected data so that analyses may span both result sets in order to find lead compounds.

We found that in the SMA screen, the main effect corrected was one in which the left and right portions of the plate had lowered luminescence intensities, probably due to evaporation from the plate edges. Each plate was analyzed, and if a row, column, or edge effect was located, it was automatically corrected using this technique. As a result of our systematic error correction technique, low-scoring compounds were slightly boosted in signal, and higher-scoring values were slightly lowered in signal (Figure 3D). Since we are looking for high-scoring compounds, this slight correction does not significantly affect the results of our assay. Even so, this technique restored a small collection of compounds that would otherwise have been missed (Figure 3D). In general, we combine the results of the corrected and noncorrected data when looking for follow-up or lead compounds. Finally, comparing histograms of the corrected and noncorrected data shows that the corrected mean value, −0.1, is closer to the expected percent enhancement of zero. The standard deviation of the corrected data is also reduced, which is also expected since removal of spatial artifacts should suppress spatially correlated outliers.

Once compounds were ranked using our scoring mechanism, results of the *SMN2*-LUC (i.e., *SMN2*-minigene-reporter*)* screen were analyzed. A self-organizing map (SOM) [41] was generated through SLIMS by mapping the structural space of our compound library (Figure 4A). SOMs have the property that regions closer to each other contain more similar compounds than regions farther away. A structural fingerprint is computed for each compound loaded into the SLIMS database. This fingerprint is binary vector, representing structures located within the compound. This vector is primarily intended to speed up substructure searching but has the side benefit of being usable to cluster compounds together based on structure. The SOM illustrates that the Comgenex portion of our library is composed of molecules in densely packed regions of descriptor space, which is likely due to the combinatorial nature of the library. The lower SOM shows how compounds that have greater than 50 percent enhancement span this structural space. As expected, *SMN2*-LUC active compounds, shown crosshatched in white, indicate that hits tend to span descriptor space and range from large stereochemically diverse structures such as aclarubicin to simpler structure compounds like chrysin. Conversely, if the active compounds had clumped together

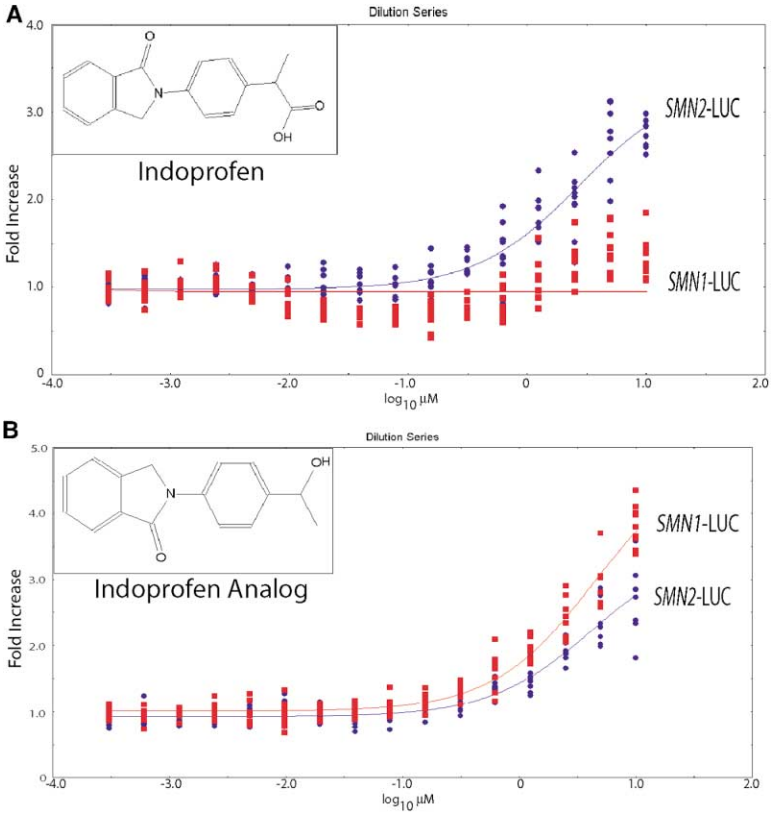

Figure 5. Analysis of Dilution Series Data in SLIMS

Dilution series for (A) indoprofen and (B) an indoprofen analog. From these analogs, only indoprofen was *SMN2* selective and had significant increases in *SMN2* production over *SMN1* production. The indoprofen analog actually had increased *SMN1* production compared to *SMN2*.

in descriptor space, then we could have simply chosen a few representative examples of structural families for follow-up testing.

In addition to exploring the chemical similarity of active compounds, SLIMS can find related analogs within tested libraries. Analogs are useful both in validating hit compounds and in making an initial assessment of the structure-activity relationship (SAR) surrounding each chemical scaffold. These scaffolds can be discovered by analyzing the SOM or by allowing SLIMS to automatically extract interesting chemical scaffolds from active compounds and searching for like compounds. Finding active analogs to lead compounds validates the leads as genuine, and not statistical outliers, although the same confidence can be gained with replicate testing. We found a collection of analogs within the *SMN2*-LUC SMA hits (Figure 4B). All of these were chosen for testing in the secondary *SMN1*-LUC (*SMN1*-minigene-reporter) screen.

Each compound chosen was tested in a 2-fold, 16-point dilution series at a maximal concentration of 10 μM in both *SMN2*-LUC and *SMN1*-LUC cell lines. We sought compounds that enhanced *SMN2* reporter production but not *SMN1* reporter production. By querying the *SMN2*-LUC and *SMN1*-LUC data, we identified compounds with good *SMN2*-LUC (*SMN2*) response and poor *SMN1*-LUC (*SMN1*) response, allowing selection of compounds based on *SMN2* selectivity.

These compounds were tested in a dilution series in both the *SMN2*-LUC and *SMN1*-LUC cell lines. SLIMS has simple wizards, similar to the standard protocol creation wizard, for rapidly creating dilution series plates electronically. When dilution plates were added to SLIMS, they were annotated with the cell type, and dilution curves were automatically calculated, fitting the resulting curve fit with the standard Hill equation. Figure 5 shows the results of the 2-fold dilution for indoprofen, a confirmed *SMN2*-selective compound, and a close analog that is not *SMN2* selective.

After the screening process was completed, the hit compounds were further analyzed to determine whether they were members of a class of active structural analogs. One particularly promising class of actives was based on 6,7-dihydroxyflavone, which was *SMN2*-LUC selective. This analysis is similar to the so-called "leader clustering," where promising lead compounds are used to analyze the assay response of compounds with related structures. SLIMS performs this by plotting the similarity score to a selected compound against the assay response. As seen in Figure 6, by computing a similarity score between 6,7-dihydroxyflavone and the compound library, a small class of actives was discovered, supporting the active and selective hypothesis. All of the compounds in this class were selected for follow-up secondary screening.

Finally, we used a built-in SLIMS module to search the Medline database to identify potential mechanisms that might be influencing *SMN2* protein production. This procedure greatly accelerates the search relative to the time it would take using a Pubmed web-based interface. We submitted the active compounds to a Medline search that identified potential mechanisms of action. Each mechanism was ranked based on a statistical analysis of compound-mechanisms pairings in Medline abstracts

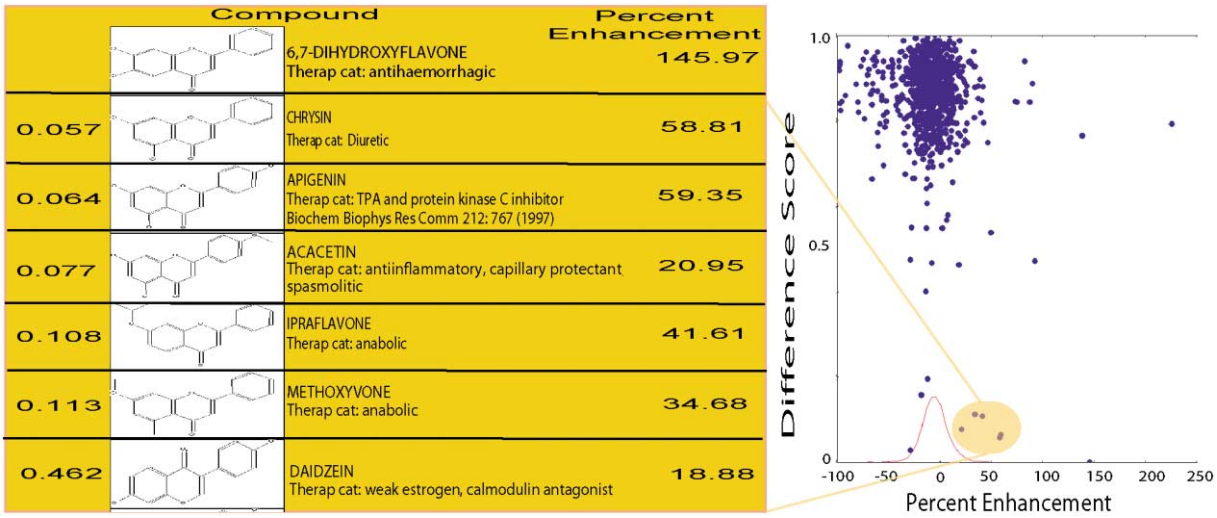

Figure 6. One of the Potential Lead Compounds, 6,7-Dihydroxyflavone, which Is Also Selective against *SMN1*-LUC in the Primary Screen, Is a Member of a Structural Family that Has Many Active Members

SLIMS generates an analysis of assay response versus similarity to the target compound. The highlighted group of compounds are similar to the potential lead compound and have a difference score close to zero. The figure plots each of the compounds' similarities to 6,7-dihydroxyflavone against their mean percent enhancement. The highlighted points are the active members of this class. The histogram in red is a histogram of the library's response to this assay and indicates that this structural class has a significant active response compared to the library as a whole.

[40]. Table 1 identifies a collection of mechanisms that might be influencing *SMN2* reporter production.

Future Directions

We are currently testing a world wide web interface to the SLIMS database. This service allows users to locate the structures of compounds on various plates in the database and to perform substructure and similarity searching to query the library as well as interface to the Lab Inventory System. The web service also provides access to results by displaying SLIMS projects and experiments and allowing users to download the specified results in either Microsoft Excel format documents or as SLIMS ready-to-open databases, as well as handling the laboratory inventory and ordering.

Table 1. Library NINDS Version 2

Mechanism	Drug	Drug
Antioxidants	aclarubicin	
Pp	apigenin	
DNA	aclarubicin	
Protein synthesis	aclarubicin	
Glycosylation	aclarubicin	
Anticoagulants	anisindione	
Cyclooxygenase	indoprofen	rhapontin
Jun	anisindione	
Antibiotics	aclarubicin	
Up	rhapontin	
Anti-inflammatory agents	indoprofen	
Cdk2	rhapontin	
Cell cycle	rhapontin	
Phosphatase	rhapontin	
Cdk6	rhapontin	
Tf	anisindione	
Antibiotic	aclarubicin	
Anticarcinogenic agents	rhapontin	
Actin	apigenin	

Medline search based on *SMN2*-LUC hits. The annotated compound library can be used to search for mechanisms that affect a cellular process. This table shows potential mechanisms of our *SMN2*-LUC active compounds.

Significance

We have created a powerful software tool for tracking and analyzing chemical screening data. This tool is particularly well suited to chemical genetic screens because it allows for flexible and dynamic analyses and collaborative sharing of data, which are hallmarks of such research projects. This software system not only tracks and analyzes chemical screening data, but it has several novel features not found in existing software tools: it automatically identifies and corrects systematic error, it automatically queries Medline for potential mechanisms of action for compounds of interest, it allows rapid and flexible protocol creation, and it allows simple transfer and sharing of data for collaboration or publication. This tool and the corresponding source code are freely available for academic use. We hope that this system will enable sophisticated chemical genetic screening efforts in a wide variety of academic laboratories.

Experimental Procedures

Cell Culture

Immortalized human cervical carcinoma cells (c33a) stably transfected with a construct that contains either *SMN1* or *SMN2* exons 6–8, neomycin resistance, and a luciferase reporter gene [39] were cultured in Dulbecco's modified Eagle's meduim (DMEM) (JRH Biosciences/#56499-10L) that was supplemented with 10% (v/v) fetal bovine serum (FBS) (Sigma/#F2442), G418 selective antibiotic

(GIBCO-BRL/#11811-031), and other antibiotics (50,000 units penicillin/L DMEM, 50 mg streptomycin/liter DMEM) (Sigma/#P4333). Cells were allowed to grow at 37°C with 5% carbon dioxide in Corning 175 cm^2 vented tissue culture flasks (VWR Scientific/#29560-970).

Type I spinal muscular atrophy (SMA)-affected human primary fibroblasts (#GM03813) and carrier parents (#GM03814, mother and #GM03815, father) were obtained live with low passage number from Coriell Cell Repositories. They were cultured in minimum essential medium (MEM) with Earle's salts and nonessential amino acids (GIBCO/#10370) that was supplemented with 15% (v/v) fetal bovine serum (FBS) (Sigma/#F2442), 2 mM l-glutamine (US Biologicals/#G7120), and antibiotics (50,000 units penicillin/liter MEM, 50 mg streptomycin/liter MEM) (Sigma/#P4333). Cells were allowed to grow at 37°C with 5% carbon dioxide in Corning 175 cm^2 vented tissue culture flasks (VWR Scientific/#29560-970).

Preparation of Compound Library for Primary Screening

Compound libraries were either obtained at a concentration of 4 mg/ml in dimethyl sulfoxide (DMSO) or were solubilized and plated at this concentration in 384-well stock "mother" plates. These compounds were then diluted into an aqueous medium to create "daughter" plates as follows: 147 μl of Dulbecco's modified Eagle's medium was dispensed, using a Zymark SciClone ALH, into each well of a Greiner 384-well, clear, polypropylene, 22 mm deep daughter plate (E&K Scientific/#EK-30202). Three microliters from the mother plate was transferred to the daughter plates using a Zymark SciClone ALH with 384-well fixed-tip pipetting head. This resulted in a compound concentration of 80 μg/ml in each well of the daughter plates.

Cell Seeding into 384-Well Plates and Compound Primary Screening

C33a cells were detached from the flasks using trypsin-EDTA (0.25% trypsin, 1 mM EDTA • 4 Na) (Life Technologies/#15050065). The cells were rinsed with 3 ml trypsin-EDTA, which was immediately aspirated. The cells were then incubated for 5–10 min at 37°C with an additional 3 ml of trypsin-EDTA. The trypsin enzyme was neutralized with 7 ml of media (normal C33a culture media lacking G418). Two to three 10 ml aliquots were combined and centrifuged at 228 × g (1000 rpm) for 5 min. The supernatant was aspirated, and the cells were resuspended at a concentration of 200,000 cells/ml in G418-free media. The cell suspension was kept at 14°C, and constant stirring at 300 rpm ensured that the suspension was maintained. Fifty-seven microliters of cell suspension was dispensed, using a Zymark SciClone ALH, into each well of a Nunc 384-well, opaque, white, tissue-culture-treated, 13 mm deep, assay plate (VWR Scientific/#62409-072) for a concentration of 11,400 cells/well. Three microliters from the daughter plate was transferred to each assay plate using a Zymark SciClone ALH with 384-well fixed-tip pipetting head; this resulted in a compound concentration of 4 μg/ml. The plate was covered with a lid and incubated at 37°C and 5% carbon dioxide for 48 hr.

SLIMS Development

SLIMS was created using many open source tools. SLIMS was programmed in Python (http://python.org/) using the wxWidgets (http://wxwidget.org/) GUI framework. The standalone database used is metakit from Equi4 software (http://equi4.com/). SLIMS is hosted on the sourceforge open source repository and can be downloaded from http://slims.sourceforge.net/. Several software toolkits were employed for SLIMS development, including scientific python (http://www.scipy.org/), used primarily for building the error correction routines.

Compound Library Creation

A portion of the compounds used in the SMA screen were selected using SLIMS. These compounds formed our TIC library. To facilitate selection of molecules for screening, SLIMS employs a molecular scripting system component developed by us. The TIC library was assembled from natural product libraries purchased from commercial sources and was designed by selecting several chemical features according to the following rules: (1) Select compounds for library from naturally derived sources. (2) Limit the number of steroids. Many natural products and compounds derived from natural products fall into the steroid class. While useful, these compounds are heavily studied and can have pleiotropic effects. (3) Maximize the number of chiral centers. The number of chiral centers in a molecule is a basic indicator of structural complexity. (4) Minimize unsaturated rings. Combinatorial libraries tend to have a large number of aromatic rings, while natural products tend to have saturated heterocyclic rings.

When this library was assembled, the FROWNS scripting system (http://frowns.sourceforge.net/), a component of SLIMS, was heavily employed during the selection process. Using this scripting system, collaborators can precisely duplicate the selection criteria on their own data sets or duplicate the creation of the TIC library.

SLIMS and all the raw data described herein are available at http://www.StockwellLab.org/slims.

Acknowledgments

This research of B.R.S. was supported in part by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, by the National Cancer Institute (R01CA97061), and by Andrew's Buddies.

Received: July 10, 2004
Revised: August 12, 2004
Accepted: August 31, 2004
Published: November 29, 2004

References

1. Schreiber, S.L. (1998). Chemical genetics resulting from a passion for synthetic organic chemistry. Bioorg. Med. Chem. *6*, 1127–1152.
2. Crews, C.M., and Splittgerber, U. (1999). Chemical genetics: exploring and controlling cellular processes with chemical probes. Trends Biochem. Sci. *24*, 317–320.
3. Stockwell, B.R. (2000). Chemical genetics: ligand-based discovery of gene function. Nat. Rev. Genet. *1*, 116–125.
4. Stockwell, B.R. (2000). Frontiers in chemical genetics. Trends Biotechnol. *18*, 449–455.
5. Koh, B., and Crews, C.M. (2002). Chemical genetics. A small molecule approach to neurobiology. Neuron *36*, 563–566.
6. Stockwell, B.R. (2002). Chemical genetic screening approaches to neurobiology. Neuron *36*, 559–562.
7. Lokey, R.S. (2003). Forward chemical genetics: progress and obstacles on the path to a new pharmacopoeia. Curr. Opin. Chem. Biol. *7*, 91–96.
8. Schreiber, S.L. (2003). The small-molecule approach to biology: Chemical genetics and diversity-oriented organic synthesis make possible the systematic exploration of biology. Chem. & Eng. 1199 News *81*, 51–61.
9. Hertzberg, R.P., and Pope, A.J. (2000). High-throughput screening: new technology for the 21st century. Curr. Opin. Chem. Biol. *4*, 445–451.
10. Jenssen, T.K., Laegreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. *28*, 21–28.
11. Rindflesch, T.C., Tanabe, L., Weinstein, J.N., and Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac. Symp. Biocomput. *2000*, 517–528.
12. Shi, L.M., Fan, Y., Lee, J.K., Waltham, M., Andrews, D.T., Scherf, U., Paull, K.D., and Weinstein, J.N. (2000). Mining and visualizing large anticancer drug discovery databases. J. Chem. Inf. Comput. Sci. *40*, 367–379.
13. Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J., Jr., Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., et al. (1997). An information-intensive approach to the molecular pharmacology of cancer. Science *275*, 343–349.
14. Brent, R. (2000). Genomic biology. Cell *100*, 169–183.
15. Engels, M.F., and Venkatarangan, P. (2001). Smart screening: approaches to efficient HTS. Curr. Opin. Drug Discov. Dev. *4*, 275–283.

16. Weinstein, J.N., and Buolamwini, J.K. (2000). Molecular targets in cancer drug discovery: cell-based profiling. Curr. Pharm. Des. *6*, 473–483.
17. Stanton, D.T., Morris, T.W., Roychoudhury, S., and Parker, C.N. (1999). Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery. J. Chem. Inf. Comput. Sci. *39*, 21–27.
18. Giuliano, K.A., and Taylor, D.L. (1998). Fluorescent-protein biosensors: new tools for drug discovery. Trends Biotechnol. *16*, 135–140.
19. Maeda, I., Kohara, Y., Yamamoto, M., and Sugimoto, A. (2001). Large-scale analysis of gene function in Caenorhabditis elegans by high-throughput RNAi. Curr. Biol. *11*, 171–176.
20. Roberge, M., Berlinck, R.G., Xu, L., Anderson, H.J., Lim, L.Y., Curman, D., Stringer, C.M., Friend, S.H., Davies, P., Vincent, I., et al. (1998). High-throughput assay for G2 checkpoint inhibitors and identification of the structurally novel compound isogranulatimide. Cancer Res. *58*, 5701–5706.
21. Silverman, L., Campbell, R., and Broach, J.R. (1998). New assay technologies for high-throughput screening. Curr. Opin. Chem. Biol. *2*, 397–403.
22. Simons, A., Dafni, N., Dotan, I., Oron, Y., and Canaani, D. (2001). Establishment of a chemical synthetic lethality screen in cultured human cells. Genome Res. *11*, 266–273.
23. Stockwell, B.R., Haggarty, S.J., and Schreiber, S.L. (1999). High-throughput screening of small molecules in miniaturized mammalian cell-based assays involving post-translational modifications. Chem. Biol. *6*, 71–83.
24. Tamura, S.Y., Bacha, P.A., Gruver, H.S., and Nutt, R.F. (2002). Data analysis of high-throughput screening results: application of multidomain clustering to the NCI anti-HIV data set. J. Med. Chem. *45*, 3082–3090.
25. Goh, C.S., Lan, N., Echols, N., Douglas, S.M., Milburn, D., Bertone, P., Xiao, R., Ma, L.C., Zheng, D., Wunderlich, Z., et al. (2003). SPINE 2: a system for collaborative structural proteomics within a federated database framework. Nucleic Acids Res. *31*, 2833–2838.
26. Koprowski, S.P., Jr., and Barrett, J.S. (2002). Data warehouse implementation with clinical pharmacokinetic/pharmacodynamic data. Int. J. Clin. Pharmacol. Ther. *40*, S14–S29.
27. McDowall, R.D. (1993). An update on laboratory information management systems. J. Pharm. Biomed. Anal. *11*, 1327–1330.
28. Turner, E., and Bolton, J. (2001). Required steps for the validation of a Laboratory Information Management System. Qual. Assur. *9*, 217–224.
29. Goodman, N., Rozen, S., Stein, L.D., and Smith, A.G. (1998). The LabBase system for data management in large scale biology research laboratories. Bioinformatics *14*, 562–574.
30. Fay, N., and Ullmann, D. (2002). Leveraging process integration in early drug discovery. Drug Discov. Today *7* (20 Suppl.), S181–S186.
31. Ausman, D.J. (2001). Screening's age of insecurity. Mod. Drug Discov. *4*, 32–34.
32. Bajorath, J. (2002). Integration of virtual and high-throughput screening. Nat. Rev. Drug Discov. *1*, 882–894.
33. Zhang, J.H., Chung, T.D., and Oldenburg, K.R. (1999). A simple statistical parameter for use in evaluation and validation of high throughput screening assays. J. Biomol. Screen. *4*, 67–73.
34. Mills, J.C., Roth, K.A., Cagan, R.L., and Gordon, J.I. (2001). DNA microarrays and beyond: completing the journey from tissue to cell. Nat. Cell Biol. *3*, E175–E178.
35. Root, D.E., Kelley, B.P., and Stockwell, B.R. (2002). Global analysis of large-scale chemical and biological experiments. Curr. Opin. Drug Discov. Dev. *5*, 355–360.
36. Root, D.E., Kelley, B.P., and Stockwell, B.R. (2003). Detecting spatial patterns in biological array experiments. J. Biomol. Screen. *8*, 393–398.
37. Li, C., and Hung Wong, W. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol *2*, research0032.1–0032.11. 10.1186/gb-2001-2-8-research0032.
38. Lorson, C.L., Hahnen, E., Androphy, E.J., and Wirth, B. (1999). A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Proc. Natl. Acad. Sci. USA *96*, 6307–6311.
39. Zhang, M.L., Lorson, C.L., Androphy, E.J., and Zhou, J. (2001). An in vivo reporter system for measuring increased inclusion of exon 7 in SMN2 mRNA: potential therapy of SMA. Gene Ther. *8*, 1532–1538.
40. Root, D.E., Flaherty, S.P., Kelley, B.P., and Stockwell, B.R. (2003). Biological mechanism profiling using an annotated compound library. Chem. Biol. *10*, 881–892.
41. Kohonen, T. (2001). Self-organizing maps. In Springer Series in Information Science, Volume 30 (Berlin: Springer).