HDAT: web-based high-throughput screening data analysis tools

To cite this article: Rong Liu et al 2013 Comput. Sci. Disc. 6 014006

View the article online for updates and enhancements.

Related content

- Investigation of an expert health monitoring system for aeronautical structures based on pattern recognition and acousto-ultrasonics
  Diego Alexander Tibaduiza-Burgos and Miguel Angel Torres-Arredondo

- Nanoinformatics workshop report: current resources, community needs and the proposal of a collaborative framework for data sharing and information integration
  Stacey L Harper, James E Hutchison, Nathan Baker et al.

- Preferential killing of cancer cells and activated human T cells using ZnO nanoparticles
  Cory Hanley, Janet Layne, Alex Punnoose et al.

Recent citations

- Leveraging the new predictive toxicology paradigm: alternative testing strategies in regulatory decision-making
  Timothy Malloy and Elizabeth Beryt

- Association rule mining of cellular responses induced by metal and metal oxide nanoparticles
  Rong Liu et al
HDAT: web-based high-throughput screening data analysis tools

Rong Liu1, Taimur Hassan1, Robert Rallo2 and Yoram Cohen3
1 California Nanosystems Institute, University of California, Los Angeles, CA 90095, USA
2 Departament d’Enginyeria Informatica i Matematiques, Universitat Rovira i Virgili, Avinguda Països Catalans 26, E-43007 Tarragona, Catalunya, Spain
3 Chemical and Biomolecular Engineering Department, University of California, Los Angeles, CA 90095, USA
E-mail: yoram@ucla.edu

Received 25 April 2013, in final form 15 September 2013
Published 9 October 2013
Computational Science & Discovery 6 (2013) 014006 (11pp)
doi:10.1088/1749-4699/6/1/014006

Abstract. The increasing utilization of high-throughput screening (HTS) in toxicity studies of engineered nano-materials (ENMs) requires tools for rapid and reliable processing and analyses of large HTS datasets. In order to meet this need, a web-based platform for HTS data analyses tools (HDAT) was developed that provides statistical methods suitable for ENM toxicity data. As a publicly available computational nanoinformatics infrastructure, HDAT provides different plate normalization methods, various HTS summarization statistics, self-organizing map (SOM)-based clustering analysis, and visualization of raw and processed data using both heat map and SOM. HDAT has been successfully used in a number of HTS studies of ENM toxicity, thereby enabling analysis of toxicity mechanisms and development of structure–activity relationships for ENM toxicity. The online approach afforded by HDAT should encourage standardization of and future advances in HTS as well as facilitate convenient inter-laboratory comparisons of HTS datasets.

1. Introduction

Nano-sized materials are increasingly utilized as common elements in modern industrial products and processes primarily due to their novel beneficial properties that arise at the nano-scale [1, 2]. At the same time, there is growing concern that engineered nano-materials (ENMs) may have adverse impacts on the environment and human health [3–12]. Given the rising public concern regarding the potential environmental impact of ENMs, efforts are mounting to assess the potential releases, toxicity and thus associated impacts of ENMs throughout their lifecycle [13–16]. In this regard, toxicity screening is critical for characterization of the potential hazard of ENMs in order to provide information essential for risk assessment and establishment of safe-use of ENMs [6, 14, 17–21]. However, the generation of toxicity data necessary to cope with the expected growth in number and diversity of ENMs is a formidable task. In order to address the above challenge, experimental approaches to high-throughput screening (HTS) of ENM toxicity [22–25] have been developed...
with the goal of providing rapid identification of toxic ENMs. The National Academy of Sciences has put forth a vision and strategy for HTS approaches recognizing the need for fast, robust and mechanistic platforms to assess multiple toxicants [22]. It is also noted that the US Environmental Protection Agency (US EPA) has implemented HTS approaches in its ToxCast program [26]. Recent advances in assessing ENM toxicity [8–12, 27–29] have demonstrated that HTS can efficiently generate large amounts of ENM toxicity data in support of the development of environmental and health regulatory policies for ENMs [14, 23].

The rapid expansion in HTS efforts for evaluation of ENM toxicity have led to the challenge faced by researchers in efficient and reliable processing and analyzing large amounts of HTS data. In this regard, it is important to recognize that ENM toxicity data, generated via HTS, usually involve high noise levels due to various uncontrolled effects [14, 23]. Although there are a number of publicly available tools for general HTS data analysis [30, 31], the range of statistical methods provided by these tools is limited and not geared specifically toward robust and reliable inferences of ENM toxicity from HTS data. Recently, certain statistical methods (e.g. strictly standardized mean difference (SSMD) [32–35]) and clustering analyses (e.g. the self-organizing map (SOM) [11]) have been identified as being particularly suitable for ENM toxicity data analyses.

In order to enable rapid analyses of ENM HTS data by the growing community of HTS researchers, there is a need for an effective set of tools that are publicly available with a range of data analysis methods and convenient data/analysis visualization tools. Accordingly, a unique set of integrated web-based (online) HTS data analysis tool (HDAT, publicly available [36]) was developed as a core component of the nanoinformatics infrastructure [14, 37–40] of UC Center for the Environmental Implications of Nanotechnology (UC CEIN). HDAT provides a large number of HTS plate normalization methods, various HTS summarization statistics, SOM [11]-based clustering analysis and visualization using both heat map and SOM. In the present contribution, the main features of HDAT are introduced, along with illustrations of HDAT usage with recently published ENM HTS toxicity data [8], demonstrating the potential of the present platform for advancing HTS studies of ENM toxicity.

2. High-throughput screening data analysis workflow

The generation, processing and interpretation of ENM HTS toxicity data [8, 9, 11, 12, 14, 41] typically follow the workflow depicted in figure 1. The first step in HTS data analysis involves plate data preparation and processing, including plate visualization, outlier removal and plate normalization. Plate visualization allows initial visual inspection of data from each HTS plate (e.g. evaluation of the consistency of sample replicates and the effectiveness of positive/negative controls) which can then guide the selection of suitable statistical methods for subsequent data analysis. Outlier removal from experimental HTS datasets may be required in order to exclude abnormal values (i.e. statistically inconsistent and thus unlikely to belong to the dataset) and thus ensure robust and reliable inference of ENM toxicity [41, 42]. HTS plate normalization is essential in order to account for plate-to-plate variability, remove systematic errors (e.g. positional effects [42]) and compare/combine data from different plates [35, 42].

In HTS experiments, replicates are commonly used in order to compensate for experimental variability [23]. Replicate measurements can significantly improve the reliability of estimates for sample bioactivity (e.g. ENM toxicity) [42]. In plate normalization, all sample wells (including replicates) are treated individually and further data processing (or summarization) may be deployed using various statistical methods (e.g. strictly standardized mean difference [32–35]). Based on summarized HTS data, hit-identification [35] can then be performed to identify samples of high bioactivity (i.e. ‘hits’) for further confirmatory tests, as well as toxicity response interpretation and modeling. Heat maps can subsequently be generated to depict sample bioactivity (summarized HTS data) in a color map for convenient visual inspection.

Once HTS data are statistically summarized, various data mining tasks [43] can be performed to extract information that is useful for ENM risk assessment and decision making. For example, SOM clustering analysis [11, 43] can group together ENMs of similar HTS bioactivity profiles, indicating that these ENMs might share common toxicity mechanisms. Activity–activity relationships, identified for different HTS toxicity assays (e.g. via association [44–46] and correlation analyses [47]), can also be used to guide experimental design (e.g. select independent HTS toxicity assays). Structure–activity relationships (SARs) [41, 48] can
also be learned from HTS toxicity data in order to predict toxicity of ENMs based on their physicochemical properties. Although activity–activity relationships and SARs are important in studies of ENM toxicity, such analyses are typically carried out separately from HDAT since they require the use of specialized machine learning tools.

3. Main features of high-throughput screening data analysis tools

3.1. Overview

HTS data analysis using HDAT proceeds by first arranging the data in a specific prescribed HTS format. The data are then uploaded for analysis and visualization. All HDAT operations are performed via a user-friendly interface shown in figure 2. A discussion of the main features of HDAT is provided in the following subsections with illustrations based on a recently published ENM HTS dataset for metal oxide nanoparticles (NPs) [8]. The dataset provided measured toxicological responses for murine myeloid (RAW 264.7) cells exposed to eight metal oxide NPs (Al₂O₃, CeO₂, CoO, Gd₂O₃, HfO₂, In₂O₃, Mn₂O₃ and Ni₂O₃) in the size range of ∼15–140 nm, over exposure concentration of 0.39–200 mg l⁻¹ (in a two-fold increase), and exposure periods of up to 24 h [8]. The HTS assays included cellular membrane permeability (by PI uptake), intracellular calcium flux (by Fluo4 fluorescence Indicator), reactive oxygen species production (by MitoSox Red fluorescence indicator) and mitochondrial membrane potential (by JC1 fluorescence indicator). A set of 384 multiwell plates (Greiner Bio-One, Monroe, NC) for different assays were used in the HTS experiments [8]. Each plate contained quadruplicates of the eight NPs at each concentration as well as two columns of negative control wells (i.e. in which cells were not exposed to NPs).

3.2. Standardized data format with flexible configuration

HTS utilizes a standardized HTS plate data format, which contains both data and plate configuration sections (figure 3). The configuration section describes where the samples and controls are arranged (see figure 3(a)), followed by data sections containing the actual HTS data (figure 3(b)). In the configuration section, samples of the same name are recognized by HDAT as replicates irrespective of their individual location in the HTS plate. Special labels ‘−1’, ‘1’ and ‘0’ are reserved for identification of negative control, positive control and ignored wells, respectively. Wells that are labeled as ‘ignored’ denote missing values, erroneous data or undesired data, and thus can be excluded from subsequent analyses. The allowance of flexible plate configuration is especially useful when HTS experiments adopt a randomized arrangement of samples and controls to reduce...
positional effects. The HDAT input file (comma delimited CSV file) contains plate data sections below their configuration section. An input file can have multiple configurations, which do not have to be of the same size and can be from different HTS experiments (figure 4). This design enables HDAT batch analysis of data from multiple HTS experiments.
Figure 4. Input file structure of HDAT.

Figure 5. Visualization of plate ‘Fluo-T2’ from the metal oxide ENMs HTS dataset. (a) Raw plate data and (b) plate with outliers removed from negative controls. In plate visualization, negative and positive control wells are identified by map cells of green and red borders, respectively. The sample wells are without borders and ignored wells are left empty. Control wells identified as outliers are labeled as ‘ignored’ and thus are excluded from subsequent statistical analysis.

3.3. Plate operations

Plate analyses offered by HDAT include plate visualization, outlier removal and plate normalization. The provided plate visualization is highly customizable and allows users to tune map cell size, color scheme and lower/upper color scale limits to best represent their plate data. As an example, figure 5(a) depicts visualization of plate ‘Fluo-T2’ (Fluo4 assay values after 2 h exposure period) for the metal oxide NP dataset described in section 3.1.

HDAT adopts the box-plot approach [43] to identify and remove outliers from plate control wells. Given a set of data points, a box plot identifies those values outside the range \([Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]\)
as outliers (in which $Q_1$ and $Q_3$ are the first and third quartiles of the data, respectively) [43]. For a normally distributed population, data points outside the above range are unlikely (<1%) members of the control population. Figure 5(b) illustrates the ‘Fluo-T2’ plate after removal of outliers in control wells. It should be noted that outlier removal is generally conducted for (positive/negative) control wells, but typically excluding sample wells due to the usually limited sample replicates in HTS experiments [42].

Subsequent to removal of outliers, a number of different plate normalization methods [31, 42, 49–52] can be used in HDAT. Here it is important to note that plate normalization methods are associated with different hypotheses about the type of data population [31, 42]. The different HDAT plate normalization methods are listed below (where $x_i$ denotes a sample value, a negative/positive control value is represented by $c_−/c_+$, and $µ$ and $σ$ denote average and standard deviation, respectively):

a. Signal to negative control ratio [42]: $µ_{xi}/µ_{c−}$ (also known as fold increase).
b. Signal to positive control ratio [42]: $µ_{xi}/µ_{c+}$.
c. Signal to noise ratio [31]: $(µ_{xi} − µ_{c−})/σ_{c−}$.
d. Normalized percent inhibition [42]: $(µ_{c+} − µ_{xi})/(µ_{c−} − µ_{c−})$.
e. Z-score [31, 42]: $(x_i − µ_x)/σ_x$.
f. Robust Z-score [42]: a robust version of Z-score using median and median absolute deviation (MAD=median($x_i$−median($x$))) in place of the average and standard deviation in Z-score.
g. Median Polish [42, 49, 50, 52]: a method to reduce the positional effects [42]. Median Polish proceeds by alternately removing the row and column medians, and continues until the proportional reduction in the sum of absolute residuals is less than a given threshold. The residual of the plate well in the $i$th row and $j$th column obtained by Median Polish is given by $r_{ij} = x_{ij} − x_{i}^{′} = x_{ij} − (µ′ + R_i + C_j)$, in which $µ′$ is the estimated average of the plate with $R_i$ and $C_j$ denoting the estimated systematic measurement offsets for the $i$th row and $j$th column, respectively.
h. B-score [31, 42, 49–52]: a robust analogue of the Z-score which is used to reduce measurement bias due to positional effects and is less sensitive to outliers. The B-score can be calculated based on Median Polish as $r_{ij}/\text{MAD}(r_{ij})$.

3.4. Analysis of high-throughput screening data accounting for replicates

HDAT provides a number of statistical methods for HTS data analysis in order to derive reliable estimation of sample bioactivity from replicated measurements. Such methods include mean (with standard deviation), median (with MAD), Z factor [31, 35, 53] and SSMD (together with its standard deviation) [32–35]. The mean and median are simple approaches to estimate sample activity from replicated measurements, while Z factor and especially SSMD are advanced statistics that consider together the mean and variance of both the sample and control. For a given sample $x$, its Z factor and SSMD are defined by $1−3(σ_x + σ_c)/|µ_x − µ_c|$ and $(µ_x − µ_c)/\sqrt{σ_x^2 + σ_c^2}$, respectively. Once the data are analyzed by one of the above approaches, the data can be visualized via customizable heat maps (see figure 6 for a heat map generated by HDAT for the metal oxide HTS data quantified via the SSMD method). Hit identification [35] can also be accomplished using HDAT for summarized data to detect samples for which there is significant up and/or down regulations for specific assays. A reduced heat map can then be generated showing only the samples identified as ‘hits’. This is illustrated in figure 7 for the ‘hits’ identified by HDAT (for the same dataset analyzed in figure 6) using a threshold of SSMD > 3, which indicates that there is a 99% certainty that the sample activity is higher than the control value, assuming that the difference between two populations is normally distributed.

3.5. Clustering analysis and visualization based on a self-organizing map

SOM [11, 54] analysis is used in HDAT for identifying clusters containing samples of similar behavior from multi-dimensional HTS data (e.g. HTS using multiple toxicity assays). SOM analysis builds an ordered two-dimensional visualization from summarized HTS data, where the most similar samples are grouped into the same SOM unit with similarity between different SOM units indicated by their geometric (Euclidean) distance
Figure 6. SSMD heat map for a metal oxide ENM HTS dataset (section 3.1). Each row represents SSMD values calculated for each HTS plate (identified by assay name and exposure period (−T##)). NP concentrations (0.39–200 mg l^{-1}) as per the fold increase specified in section 3.1 are identified by the numbers (01–10) appended to the NP names.

Figure 7. Reduced heat map showing ‘hits’ (denoted by red cells in the map) identified from the detailed heat map shown in figure 6.

such that the original distance relationships (topology) are preserved [11, 54, 55]. SOM units are further grouped together into different clusters representing major groups of samples of similar behavior. As an illustration, figure 8 depicts an SOM built by HDAT from the SSMD summarized data of the metal oxide ENM dataset described in section 3.1. In this example, Mn_2O_3, CoO and Ni_2O_3 ENMs, which have been identified as toxic (see figure 7) at concentrations above about 50 mg l^{-1}, are clustered in the bottom cluster, while ENMs exhibiting lower toxicity at lower concentrations are clustered in the right-hand cluster of the SOM. There are various ways of visualizing the SSMD data via SOM and such clustering, along with an assay-specific SSMD pattern, can aid in interpretation of toxicity mechanisms [11].
Figure 8. SOM clustering based on the SSMD of the metal oxide ENM dataset (section 3.1). Five clusters (marked in different colors) of similar NPs (in terms of the HTS profiles including the four assays over the exposure period) are identified in this SOM analysis.

3.6. User assistance

HDAT provides online instructions, video demos of basic operations and specification for data formatting and uploading (see figure 1). In addition, a sample ENM HTS toxicity dataset (i.e. the metal oxide NP toxicity dataset used in the present work to illustrate HDAT) is provided to allow users to explore and experiment with various HDAT functions. Another important feature built into HDAT is a discussion forum to support the growing community of HTS researchers.
4. Applications and merits of high-throughput screening data analysis tools

HDAT has been built as the HTS data analysis platform of the UC CEIN. HDAT has enabled rapid analysis of the growing HTS database of ENM toxicity in support of nano-toxicological studies. UC CEIN’s HTS studies which have utilized HDAT for processing HTS data have provided HTS-based hazard ranking of NPs (metal/oxide and quantum dots) [12], evaluation of toxicity-related cell signaling pathways for metal/oxide NPs [11], assessment of the difference in response of undifferentiated and differentiated primary human bronchial epithelial cells to cationic mesoporous silica NPs (coated with polyethyleneimine) [9], investigation of cellular oxidative stress induced by metal oxide NPs [8] and nano-SAR development for metal oxide NPs [41]. HDAT is available for public use, has generated widespread interest in the nanoinformatics community and is being used by a number of government agencies and academic institutions for HTS data analysis. Over the year since its initial release, there have been over 20 000 instances of HDAT utilization by users from 20+ countries, indicating the importance of a publicly available online HTS data analysis tool for the growing nanoinformatics community. Although developed primarily for analysis of HTS ENM data, HDAT can be used for other applications data that can be cast in a matrix format.

Acknowledgments

This material is based upon work supported by the National Science Foundation and the Environmental Protection Agency under Cooperative Agreement Number DBI-0830117. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Environmental Protection Agency. This work has not been subjected to EPA review and no official endorsement should be inferred. RR also acknowledges support provided by the Spanish Government (Project CTM2011-24303), Generalitat de Catalunya (2009SGR-01529) and the EU Commission (MODERN, contract no. 309314).

References

[18] Linkov I et al 2011 A decision-directed approach for prioritizing research into the impact of nanomaterials on the environment and human health Nature Nanotechnol. 6 784–7
[23] Damoiseaux R et al 2011 No time to lose—high throughput screening to assess nanomaterial safety Nanoscale 3 1345–60
[29] Liu R et al 2012 Automated phenotype recognition for zebrafish embryo based in vivo high throughput toxicity screening of engineered nano-materials PLoS ONE 7 e35014
[31] Ling X F B 2008 High throughput screening informatics Comb. Chem. High Throughput Screening 11 249–57
[33] Zhang X D 2007 A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays Genomics 89 552–61
[34] Zhang X D 2007 A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays J. Biomolecular Screening 12 645–55
[36] High Throughput Screening Data Analysis Tools 2011 Available from nanoinfo.cein.ucla.edu/public/hdat
[37] Nanoinformatics 2020 Roadmap Available from eprints.internano.org/607/
[38] Thomas D G et al 2011 Informatics and standards for nanomedicine technology WIREs Nanomed. Nanobiotechnol. 3 511–32
[41] Liu R et al 2011 Classification nanoSAR development for cytotoxicity of metal oxide nanoparticles Small 7 1118–26


