ELSEVIER

Alzheimer's & Dementia ■ (2018) 1-15

Alzheimer's & Dementia

Featured Article

# Intelligent and effective informatic deconvolution of "Big Data" and its future impact on the quantitative nature of neurodegenerative disease therapy

Stuart Maudsley[a,b,*,1], Viswanath Devanarayan[c,1], Bronwen Martin[a], Hugo Geerts[d], on behalf of the Brain Health Modeling Initiative (BHMI)

[a]*Department of Biomedical Research, University of Antwerp, Antwerp, Belgium*
[b]*VIB Center for Molecular Neurology, Antwerp, Belgium*
[c]*Exploratory Statistics, AbbVie, Souderton, PA, USA*
[d]*In Silico Biosciences, Inc., Berwyn, PA, USA*

**Abstract** Biomedical data sets are becoming increasingly larger and a plethora of high-dimensionality data sets ("Big Data") are now freely accessible for neurodegenerative diseases, such as Alzheimer's disease. It is thus important that new informatic analysis platforms are developed that allow the organization and interrogation of Big Data resources into a rational and actionable mechanism for advanced therapeutic development. This will entail the generation of systems and tools that allow the cross-platform correlation between data sets of distinct types, for example, transcriptomic, proteomic, and metabolomic. Here, we provide a comprehensive overview of the latest strategies, including latent semantic analytics, topological data investigation, and deep learning techniques that will drive the future development of diagnostic and therapeutic applications for Alzheimer's disease. We contend that diverse informatic "Big Data" platforms should be synergistically designed with more advanced chemical/drug and cellular/tissue-based phenotypic analytical predictive models to assist in either de novo drug design or effective drug repurposing.


*Keywords:* Big data; Informatics; High-dimensionality; Alzheimer's disease; Aging; Molecular signature; Transcriptomics; Metabolomics; Proteomics; Genomics

## 1. Introduction

We are currently in the era of facile Big Data generation: vast genomic, transcriptomic, proteomic, and metabolomic data sets are generated on a daily basis [1–4]. The technological and intellectual developments that have fueled this data explosion are now becoming embedded in many drug discovery pipelines. As is often the case with emerging technologies, there is first a wave of data-generating platforms and then a second wave of analytical systems that are created to aid with the interpretation and exploitation of these data. Presently, the collection and processing time for high-dimensionality data is becoming problematical, thus it is currently difficult to rapidly and efficiently generate a quantitative and therapeutically meaningful output of such data. High-dimensionality research in the future is likely to be dominated by the creation of intelligent analytical systems that are capable of generating effective disease diagnostic and drug development deliverables. Workflows for analyzing complex multivariate data are well documented in fields such as computer science; however, relatively fewer advances have been made in the biomedical field to condense data vectors (that exist beyond the realm of physical space) into an easily interpretable, esthetic,

[1]Equal contribution of authors.
*Corresponding author. Tel.: +32 32651057; Fax: ■■■.
E-mail address: stuart.maudsley@molgen.vib-ua.be

https://doi.org/10.1016/j.jalz.2018.01.014
1552-5260/

and translationally relevant form. To adopt a "complexity science" approach for the investigation of dementia and other brain disorders, we will need to assemble/organize Big Data resources in a rational and actionable manner. This will entail the generation of systems that allow cross-platform correlation between big data sets of distinct types, for example, transcriptomic, proteomic, and metabolomic [5–7]. Big Data analytics covers a vast computational space, ranging from bottom-up dynamical system modeling to top-down probabilistic causal approaches. Across many fields of science and physics, modeling and simulation have come to complement theory and experimentation as a key component of the scientific method.

To effectively use informatic platforms to investigate neurodegenerative diseases from a quantitative therapeutic perspective, it is first necessary to ensure that the data are synergistically curated, filtered, normalized, and extracted. To contend with the huge data analysis volumes in this process, crowdsourced curation (e.g., Community Research Education and Engagement for Data Science/GENE Expression and Enrichment Vector Analyzer) activities are potentially an important future solution for this. Second, for informative cross-domain (genomic, epigenomic, transcriptomic, RNA-sequencing, proteomic, PTM-proteomic, metabolomics) analytics that assimilate data and allow insight across diverse data streams, novel data label-free systems, such as topological data analysis (TDA), will likely be important. Elucidation of therapeutic/disease molecular signatures from diverse data streams will likely only generate informative new therapeutic strategies, once we can implement a postanalytic result integration system that is able to create a realistic "gestalt" appreciation of super-complex data. It is also imperative for effective drug design that this gestalt appreciation is based more on the quantitative, rather than the qualitative, nature of these nuanced "drug-to-data" relationships. In this review, we will provide a focused description of the latest informatic methods/platforms that can be used for disease diagnostics (e.g., for Alzheimer's disease [AD]) and quantitative drug development: Latent Semantic Analyses (Section 2.2); drug activity characterization (Section 2.2.1); electronic medical data (EMD) file analytics (Section 2.2.2); data visualization and topological methods (Section 3); network/graph and Game Theory implementation (Section 4); machine learning and pattern recognition for biomarker analyses (Section 5.1), biomedical image analysis (Section 5.2), and proteomic and mass spectrometric (MS) data analysis (Section 5.3).

## 2. Mechanisms for effective data retrieval, investigation, and therapeutic implementation

### 2.1. Curated analysis: classical pathways/ontologies applied in $n^{th}$ dimensions

At the most basic level, literature-based interrogation of individual genes/proteins from the primary data still represents an effective tool for analysis. At a "human" level of capacity, however, such an approach may ignore multiple data correlations due to clear problems with information retrieval and recall. Compared to simplistic data streams of two decades ago, high-dimensionality data workflows have revealed that cell signaling pathways, a prime target for drug development, represent only a minimal fraction of a more complex signaling network. Hence gene/protein interactions are more complex and numerous than once thought and possibly exist in an innumerable (n) number of data dimensions. Given this transition in the appreciation of biological complexity, it is clear that a "gestalt" data set appreciation is likely to yield a more accurate appraisal of signaling/therapeutic activity in a disease setting.

Linkage of biological similarities between multiple data set elements and predicted functional sequelae can be explored with informatic term association techniques, for example, gene ontology, signaling pathway analysis (Kyoto Encyclopedia of Genes and Genomes, Reactome pathways), biophysical parameters (protein family database, Protein Information Resource_Superfamilies, Simple Modular Architecture Research Tool), interactomic profiles (Biological General Repository for Interaction Datasets, Biomolecular Interaction Network Database, Molecular INTeraction database, Search Tool for the Retrieval of Interacting Genes/Proteins database), and functional effect collections/experimental molecular signatures (Molecular Signatures Database–Gene Set Enrichment Analysis, Library of Integrated Network-Based Cellular Signatures 1000 data). While individual applications can generate important results, the coherent combination of such "basic" tools underpins the future value of these canonical platforms. For example, Li et al. [8] recently demonstrated how combinatorial informatics could elucidate pathological AD mechanisms from large-scale high-dimensionality data. Li et al. [8] used a meta-analysis approach to identify differentially expressed genes in published data sets comprising 450 late-onset AD brains and 212 controls. With this large-scale data approach, using Ingenuity Pathway Analysis–based pathway analysis, Gene Set Enrichment Analysis, and PPI investigation via a PPI-network generated from Human Protein Reference Database, these researchers were able to assemble a novel large-scale late-onset AD–related data set of 3124 differentially expressed genes. Pathway analysis of these data identified several crucial late-onset AD driving processes, including nitric oxide and reactive oxygen species generation in macrophages, Nuclear factor kappa-light-chain-enhancer of activated B cells modulation and mitochondrial dysfunction [8]. Functional integration of standard informatic approaches using large-scale genomic data sets has also been used to identify potential new drug targets. Elkahloun et al. [9] used a combination of Kyoto Encyclopedia of Genes and Genomes, Ingenuity Pathway Analysis, Gene Set Enrichment Analysis, and Gene Expression Omnibus (GEO) data analytics to prioritize the angiotensin receptor II antagonist candesartan as a