BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Review

Volume 9(12)

New tools for the investigations of Neuro-AIDS at a molecular level: The potential role of data-mining

Bruno Orlando[1, 2], Luca Giacomelli[3], Francesco Chiappelli[4]* & André Barkhordarian[4]

[1]Laboratories of Biophysics and Nanobiotechnology, Department of Medical Science, University of Genova, Italy; [2]Department of Surgery, University of Pisa, Italy; [3]Free researcher, Milan, Italy; [4]UCLA School of Dentistry, CHS63-090; Los Angeles CA, 90095-1668; Francesco Chiappelli – Email: fchiappelli@dentistry.ucla.edu; Phone: 310-794-6625; Fax: 310-794-7134; *Corresponding author

Received April 24, 2013; Accepted April 30, 2013; Published July 12, 2013

Abstract:
Cognitive impairment represents the most significant and devastating neurological complication associated with HIV infection. Despite recent advances in our knowledge of the clinical features, pathogenesis, and molecular aspects of HIV-related dementia, current diagnostic strategies are associated with significant limitations. It has been suggested that the use of some biomarkers may assist researchers and clinicians in predicting the onset of the disease process and in evaluating the effects of new therapies. However, the large number of chemicals and metabolic pathways involved in the pathogenesis of neurodegeneration, warrants the development of novel approaches to integrate this huge amount of data. The contribution of theoretical disciplines, such as bioinformatics and data-mining, may be useful for testing new hypotheses in diagnosis and patient-centered treatment interventions.

Keywords: Data Mining, Computational Biology, HIV, Dementia, Translational effectiveness.

Molecular mechanisms of cognitive impairment in patients with HIV infection:

HIV encephalitis (HIVE) is the common pathologic correlate of cognitive impairment in patients with HIV infection. In the central nervous system (CNS) of these patients, brain mononuclear phagocytes are reservoirs for persistent viral infection. HIV-infected monocytes/macrophages release several viral proteins, some of which activate glial cells such as microglia and astrocytes to release chemokines, cytokines and a number of soluble neurotoxic substances. Neurotoxins, in conjunction with secreted HIV proteins, damage the synapto-dendritic axis of neurons, resulting in neuronal dysfunction and cell loss via apoptosis **[1].** Previous studies already reported that HIV proteins abnormally activate the CDK5 **[2]** and GSK3β **[3]** cascades. The deregulation of the GSK3β enzyme could contribute to HIV-induced neuronal apoptosis. In addition, HIV proteins might activate the CDK5 pathway and subsequently upset the various actions that CDK5 regulates, including synapse formation and plasticity and neurogenesis. Past studies also speculated that in patients with cognitive impairment, the neurodegenerative process might correlate with an altered expression of neurotrophic factors such as vascular endothelial growth factor (VEGF), interleukin-8 (IL-8), and fibroblast growth factors (FGFs). FGFs exert their effects via receptor tyrosine kinases, leading to inactivation of GSK3β through phosphorylation of a serine residue. Other growth factors, such as insulin growth factor-1 (IGF-I), epidermal growth factor (EGF) and platelet-derived growth factor (PDGF), causes a similar inhibition of GSK3β activity by inducing phosphorylation **[2].**

CDK5 is a protein kinase with postmitotic activity that phosphorylates cytoskeletal proteins (MAP1b, tau, NF, nestin, DCX), synaptic proteins (PSD95, synapsin, cadherin) and transcription factors (MEF2). Its activity is regulated primarily by the metabolism of the activating proteins p35 and p39. A recent investigation indicates that p25, a truncated form of p35, accumulates in neurons of patients with neurodegenerative diseases. Binding of p25 to CDK5 constitutively activates

CDK5, changes its cellular location and alters its substrate specificity. Expression of the p25/CDK5 complex results in abnormal phosphorylation of toxic substrates that induces cytoskeletal disruption, morphological degeneration and apoptosis **[4].**

Understanding the molecular pathology of cognitive impairment in patients with HIV infection: current concerns and the need of new investigation tools:
As described above, several inflammatory molecules including cytokines, chemokines, growth factors, and excitatory compounds are associated with brain inflammation and damage. Genomics and proteomics could be applied to reach a deeper understanding of the molecular mechanisms underlying complex multifactorial disorders. However, mass-scale molecular genomics and proteomics suffer from some pitfalls: gene and protein expression is not significant per se, but only if inserted in a detailed cross-talk of molecular pathways and gene/gene, gene/protein and protein/protein interactions. Many diseases, including HIVE, are complex, polygenic disorders. In these diseases, the etiology is not attributed to the expression of a single gene or to the encoded protein, but it is spread over several different genes, each having a modest effect. However, although each gene has a small effect, the overall effect of all the genes and of all the encoded proteins involved may be substantial. Therefore, the pathophysiology of complex diseases is characterized by the involvement of various biologic pathways. A simple variation in expression of a single gene or of the encoded protein is not meaningful per se, but only if put in a proper framework of interactions (i.e. physical interaction of different molecules, involvement in the same metabolic pathway, and co-expression in microarray studies). The analysis of the complex network of connections between genes/proteins may allow the identification of potential molecular markers and targets **[5]**.

The study of interaction networks requires the systematization and the analysis of a huge amount of information emerging from experimental studies, but a complete experimental analysis of all the molecules involved in a given process, including both genes/proteins and small molecules, appears a challenging task. For instance, the greatest part of genes displayed on an array is often not directly involved in the cellular process being studied. Commercial arrays with a lower number of genes – usually 150–200 – are currently available, but the genes displayed are usually once again chosen without a precise consideration of the particular target of the study **[6].** The contribution of theoretical disciplines, such as bioinformatics and data-mining, is therefore required to integrate this huge amount of data. Bioinformatics can become an added value in this context **[6].** This discipline is defined as the application of information technology to the field of molecular biology, via the development of original algorithms **[7].** Another discipline playing an important role in the analysis of genomics/proteomics experiments is data mining, i.e. extracting patterns from data, thus developing new information from previous knowledge **[8].** With these methods, a further simplification of complex information emerging from genomics/proteomics experiments becomes possible. Properly combined with bioinformatics techniques and algorithms, data mining may allow to draw a simpler, but at the same time powerful picture of complex amount of data.

Novel approaches for the analysis of the molecular events underlying neurodegeneration in HIV-infected patients:
Bioinformatics and data-mining can play a central role in the analysis and interpretation of genomic and proteomic data. In particular, these disciplines may be useful to further clarify the pathophysiology of complex diseases, which are characterized by various biologic pathways, dependent upon the contribution of a large number of genes forming complex networks of interactions **[7, 8].** We (FC in Shapshak et al, 2006) provided a preliminary description of the molecular mechanisms underlying these processes using expression and gene annotation data **[9].** This information has been stored in a user friendly, online database that, when made available in the public domain, will enable scientists to retrieve biochemical and physiological information on neuro-degeneration in HIV-infected patients. They are also developing methods to structure anatomical information (location where specific biochemical reactions occur) so that essential structural features are included along with the topological attributes **[9].**

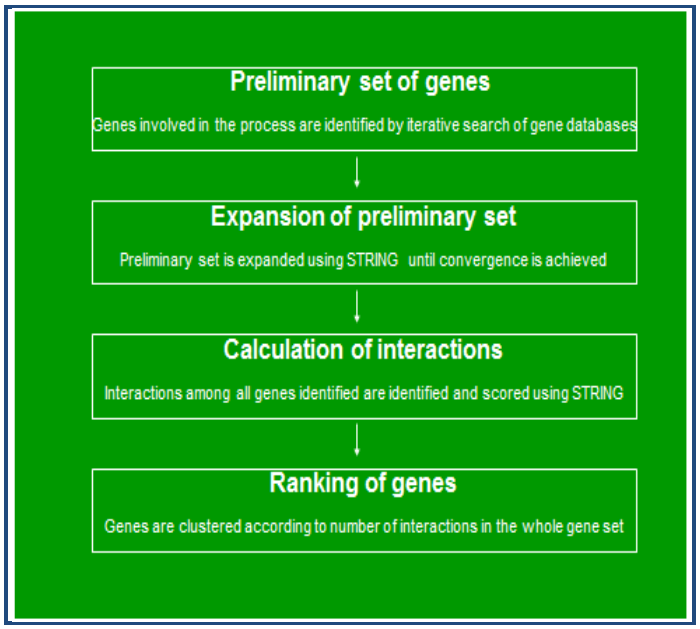

Figure 1: Flow chart of the leader gene approach.

Recently, another data-mining method, defined as the "Leader Gene approach" have been proposed **(Figure 1) [6, 10].** This mining algorithm is based on the systematic search for the genes involved in a given process; the interaction among these genes are then calculated and genes are ranked according to the number and the confidence of all experimentally established interactions, as derived from free Web-available databases, such as STRING (Search Tool for the Retrieval of Interacting Genes, Heidelberg, Germany). Genes in the highest rank are defined as ''leader genes'', since they can be preliminarily supposed to play an important role in the analyzed process. These genes may become potential targets for a targeted experimentation, which may be simpler than mass-scale molecular genomics and, at the same time, powerful **[5, 6, 10].** The Leader Gene approach was applied to different cellular processes and pathological conditions, such as the human T lymphocyte cell cycle, human kidney transplant and periodontitis **[5, 6, 10];** the results were integrated with a

targeted experimental analysis, to draw an overall picture of these processes.

First, a preliminarily set of genes with an established role in a specific process is identified by an iterative search of large-scale gene databases (PubMed, GeneBank, Geneatlas, Genecards), using several keyword-based searches and the official HGNC nomenclature. Second, the preliminary set of genes is expanded using the STRING database, excluding text-mining-derived interactions, to identify genes linked to those playing an established role in the process under study, and therefore potentially involved in it. Only interactions with a confidence score >0.9, as given by STRING, are considered. Results are then filtered to discard false positives via a keyword-based query in PubMed, until no new genes are retrieved. Third, the interactions between all the genes identified are mapped using STRING. This database gives a combined association score to each gene-gene interaction, representing its strength. The combined association scores referring to each single gene are then summed to obtain a weighted number of links (WNL). Fourth, genes are clustered (hierarchical and K-means algorithms) according to their WNL. The genes belonging to the highest class are defined as leader genes. This analysis could further confirm that data-mining of existing databases may represent a starting point to improve our knowledge about cellular processes and diseases, to formulate new hypotheses and to plan targeted experimentation **[6].** In particular, the analysis of gene interaction maps and the ranking of genes according to their interconnections might help in identifying new targets for dedicated experimental analyses, which may confirm or discard each hypothesis and suggest potential risk factors and therapy targets **[6].** Taken together, and in the current context of translational effectiveness, the utilization of the best available data in specific clinical settings for patient-centered care, the cutting-edge concepts we have discussed here confirm that bioinformatics and data-mining will have an increasingly important role in the integration of translational research findings for evidence-based diagnosis and prognostic interventions for patients with Neuro-AIDS, which require new and improved bioinformation dissemination structures, such as those we have proposed previously **[11-13].**

Acknowledgment:
We thank Professor Shapshak, and several of the stakeholders of the Evidence-Based Decision Practice-Based Research Network for the many discussions and feedback. Funded in part by Fulbright Specialist program to FC. The authors declare no conflicts of interest.

References:

[1] Minagar A *et al. Mol Diagn Ther.* 2008 **12**: 25 [PMID: 18288880]
[2] Crews L *et al. Int J Mol Sci.* 2009 **10**: 1045 [PMID: 19399237]
[3] Tong N *et al. Eur J Neurosci.* 2001 **13**: 1913 [PMID: 11403684]
[4] Shelton SB & Johnson GV, *J Neurochem.* 2004 **88**: 1313 [PMID: 15009631]
[5] Giacomelli L *et al. AMIA Annu Symp Proc.* 2007: 963. [PMID: 18694063]
[6] Giacomelli L & Nicolini C, *J Cell Biochem.* 2006 **99**: 1326 [PMID: 16795054]
[7] Kuo WP, *Adv Dent Res.* 2003. **17**: 89 [PMID: 15126216]
[8] Glasgow J *et al. Pac Symp Biocomput.* 2000 **12**: 366 [PMID: 10902184]
[9] Shapshak P *et al. Bioinformation.* 2006 **1**: 86 [PMCID: PMC1891660]
[10] Covani U *et al. J Periodontol.* 2008 **79**: 1983 [PMID: 18834254]
[11] Barkhordarian A *et al. Bioinformation* 2011 **7**: 315 [PMCID: PMC3280503]
[12] Barkhordarian A *et al. Bioinformation*. 2012 **8**: 293 [PMCID: PMC3338971]
[13] Chiappelli F *et al. Bioinformation*. 2012 **8**: 691 [PMCID: PMC3449364]

Edited by P Kangueane
Citation: Orlando *et al.* Bioinformation 9(12): 656-658 (2013)