Journal of Biomedical Informatics 54 (2015) 10–38

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Adapting simultaneous analysis phylogenomic techniques to study complex disease gene relationships

Joseph D. Romano [a,1], William G. Tharp [b], Indra Neil Sarkar [a,c,d,*]

[a] Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT 05405, USA
[b] Department of Medicine, Endocrinology Unit, University of Vermont, Burlington, VT 05405, USA
[c] Center for Clinical and Translational Science, University of Vermont, Burlington, VT 05405, USA
[d] Department of Computer Science, University of Vermont, Burlington, VT 05405, USA

ARTICLE INFO

*Article history:*
Received 15 May 2014
Accepted 5 January 2015
Available online 13 January 2015

*Keywords:*
Bioinformatics
Phylogenetics
Simultaneous analysis
Alzheimer Disease
Comparative genomics

ABSTRACT

The characterization of complex diseases remains a great challenge for biomedical researchers due to the myriad interactions of genetic and environmental factors. Network medicine approaches strive to accommodate these factors holistically. Phylogenomic techniques that can leverage available genomic data may provide an evolutionary perspective that may elucidate knowledge for gene networks of complex diseases and provide another source of information for network medicine approaches. Here, an automated method is presented that leverages publicly available genomic data and phylogenomic techniques, resulting in a gene network. The potential of approach is demonstrated based on a case study of nine genes associated with Alzheimer Disease, a complex neurodegenerative syndrome.

The developed technique, which is incorporated into an update to a previously described Perl script called "ASAP," was implemented through a suite of Ruby scripts entitled "ASAP2," first compiles a list of sequence-similarity based orthologues using PSI-BLAST and a recursive NCBI BLAST+ search strategy, then constructs maximum parsimony phylogenetic trees for each set of nucleotide and protein sequences, and calculates phylogenetic metrics (Incongruence Length Difference between orthologue sets, partitioned Bremer support values, combined branch scores, and Robinson–Foulds distance) to provide an empirical assessment of evolutionary conservation within a given genetic network. In addition to the individual phylogenetic metrics, ASAP2 provides results in a way that can be used to generate a gene network that represents evolutionary similarity based on topological similarity (the Robinson–Foulds distance).

The results of this study demonstrate the potential for using phylogenomic approaches that enable the study of multiple genes simultaneously to provide insights about potential gene relationships that can be studied within a network medicine framework that may not have been apparent using traditional, single-gene methods. Furthermore, the results provide an initial integrated evolutionary history of an Alzheimer Disease gene network and identify potentially important co-evolutionary clustering that may warrant further investigation.

## 1. Introduction

Classical genetic diseases typically arise due to isolated genetic changes within a single gene or allele [1]. Many of these "simple" or "monogenic" diseases follow Mendelian patterns of inheritance. The responsible genetic lesion is often the result of an insertion or deletion event, or the transversion/transposition of a nucleotide. The probability for transmission of simple genetic disorders may thus be easily predicted and generally follow sex-linked or autosomal patterns of heredity. Classic examples of monogenic disorders include cystic fibrosis, sickle cell anemia, and achondroplasia [2–4]. By contrast, complex diseases or disorders may not follow clear hereditary patterns or be diagnosed based on isolated genetic lesions. However, many complex diseases such as cardiovascular disease, type 2 diabetes mellitus, and Alzheimer Disease occur with higher frequency among families and close genetic relatives – suggesting that the interaction of genetic elements may play a central

[*] Corresponding author at: Center for Clinical and Translational Science, University of Vermont, 89 Beaumont Avenue, Given Courtyard S350, Burlington, VT 05405, USA. Fax: +1 802 656 4589.

*E-mail address:* neil.sarkar@uvm.edu (I.N. Sarkar).

[1] Current address: Department of Biomedical Informatics, Columbia University Medical Center, 622 W 168th Street, Presbyterian Hospital, 20th Floor, New York, NY 10032, USA.

http://dx.doi.org/10.1016/j.jbi.2015.01.002
1532-0464/© 2015 Elsevier Inc. All rights reserved.

role in their pathogenesis, beyond environmental or behavioral factors [5]. Identifying risks for complex diseases and developing new approaches for treating or preventing them may benefit from high-throughput, computational, or bioinformatics based approaches. Related advances in biotechnology have facilitated the identification of genotypes that may be factors involved in the heritability of complex genetic diseases [6]. For example, specific genotypes can be associated with a probabilistic value of susceptibility relative to the gene(s) they influence and thus correlated with a disease phenotype [1,7–9].

Due to limited knowledge about the specific mechanisms by which multiple genetic factors may influence complex diseases, pharmacotherapies are often aimed at managing symptoms or laboratory values, and are therefore reactionary and not preventative. Thus, the approach to complex disease management necessarily extends beyond pharmacotherapy, attempting environmental and behavioral changes through patient education or lifestyle modification [2,10]. A major current goal of biomedical research is therefore to better characterize complex relationships between contributing factors associated with complex diseases for identifying possible targets for therapeutic intervention. Genetic background influences the susceptibility to complex disease, which is an artifact of the structural or functional relationships between some or all members of a disease gene network [3,7]. These relationships may include direct physical interaction between the protein products of the genes, parallel functionality in metabolic pathways, or co-localization of protein products in a certain cell or tissue type [4,7]. These data are not easily elucidated using approaches that are focused on a single gene or pathway, and instead require a broader systems-based methodology. Understanding the shared history of multiple genes may provide guidance in developing approaches that target multiple genes that have evolved to work together through evolutionary time.

Complicating the assessment of such systems-based methodologies is the lack of reference standards for benchmarking approaches for discovering how myriad complex disease genes work in the context of disease phenotypes. It should therefore be noted that methodologies for interpreting multiple genes associated with complex disease may not quantifiably be benchmarked against previously used methods that focus on single-gene analysis. Instead, methodologies for multiple gene analysis can be seen as identifying *potential* relationships between genes, necessitating the development of benchmarks and controls that can be performed internally against known genetic interactions (and cases where one can be fairly certain that no interaction is taking place).

Within the context of translational bioinformatics, there have been a limited number of attempts to address understanding complex diseases that accommodate multiple disease genes simultaneously. One method has described the use of Mendelian genetic traits that occur in coincidence with complex genetic diseases to predict underlying mechanisms of those diseases, primarily by linking diverse biomedical databases [11]. Another approach involves the linking of complex disease genes to other diseases based on molecular similarity, which adapts a vector space model approach [12].

Perhaps the most significant approach developed to date for studying the potential impact of multiple disease genes is that of "network medicine," which systematically describes the rise of disease phenotypes as perturbations in the normal interactions of molecular, environmental, and population networks rather than a single macromolecule or biological pathway [13]. Network medicine postulates relationships between genes based on observed interacting phenomena, accounting for numerous genetic events or environmental factors may contribute to similar phenotypic results in one class of organism but not in another. Notably missing from network medicine approaches to data is the inclusion of models that reflect evolutionary similarities between genes that comprise a disease network. Evolutionary models may provide a perspective to the network medicine approach and enable the inclusion of ancient environmental and population data into the construction of a disease network and may identify both previously unknown factors contributing to disease development and new model organisms for understanding disease pathology.

Phylogenetic analyses infer potential evolutionary relationships based on similarities implying common descent from shared ancestry and are performed on data sets consisting of physical, functional, or molecular representations [14]. Genomic analyses typically construct the analytic matrix using nucleotide or amino-acid sequences from different individuals or species (termed "taxa"; singular "taxon"). Classically, the resulting data are presented as trees where the branching points (termed "nodes") give rise to hierarchical groupings of more similar taxa (akin to leaves on a branch). These trees can be used to explore potential patterns of divergence from a common ancestor as well as the

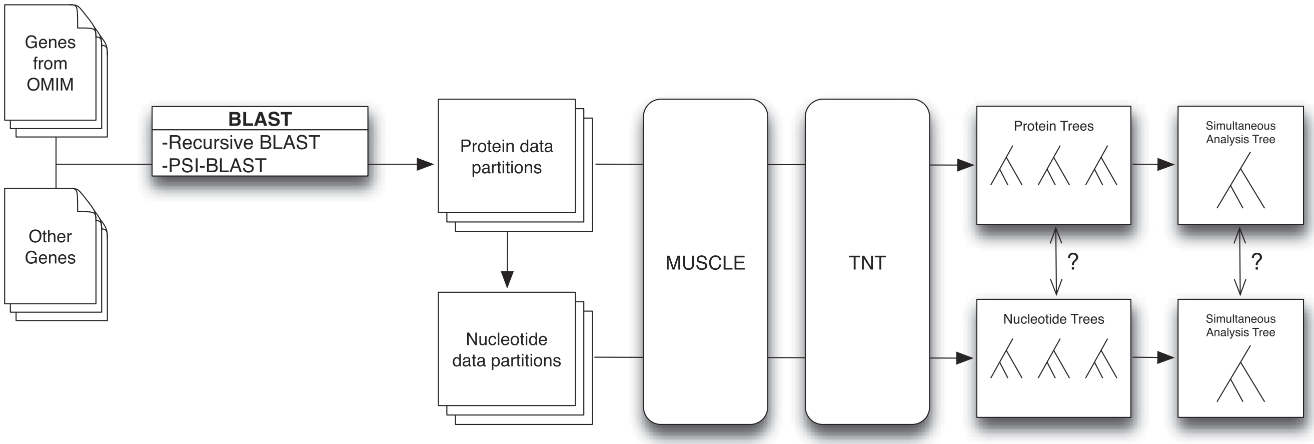

**Fig. 1.** Overview of ASAP2 workflow. The process, as implemented in the study, begins with the providing of GenBank IDs for protein sequences, which may originate from reference resources like OMIM or other user chosen sequences. A combination of a highly specific recursive BLAST+ approach and PSI-BLAST is used to identify sequence-similarity based orthologues (using a stringent *E*-value cutoff of 0.0). For each orthologue protein sequence identified, its corresponding nucleotide GenBank entry is retrieved based on metadata within the protein GenBank sequence. The remaining workflow follows the standard process for Simultaneous Analysis (SA) for both the protein and nucleotide sequence sets (called "partitions" in SA): Sequence Alignment (e.g., using MUSCLE) and phylogenetic tree building (e.g., using TNT). The resulting trees are then compared for each protein and nucleotide partition as well as for the overall protein or nucleotide SA tree. MUSCLE: Multiple Sequence Comparison by Log-Expectation (Multiple sequence alignment software). TNT: Tree analysis using New Technology (Maximum Parsimony phylogenetic analysis software).