Published in final edited form as:

Proteomics Clin Appl. 2008 February ; 2(2): 122–134.

Evolutionary medicine: A meaningful connection between omics, disease, and treatment

Mones Abu-Asab¹, Mohamed Chaouchi², and Hakima Amri^{2,*}

¹Laboratory of Pathology, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

²Department of Physiology and Biophysics, School of Medicine, Georgetown University, Washington, DC, USA

Abstract

The evolutionary nature of diseases requires that their omics be analyzed by evolution-compatible analytical tools such as parsimony phylogenetics in order to reveal common mutations and pathways' modifications. Since the heterogeneity of the omics data renders some analytical tools such as phenetic clustering and Bayesian likelihood inefficient, a parsimony phylogenetic paradigm seems to connect between the omics and medicine. It offers a seamless, dynamic, predictive, and multidimensional analytical approach that reveals biological classes, and disease ontogenies; its analysis can be translated into practice for early detection, diagnosis, biomarker identification, prognosis, and assessment of treatment. Parsimony phylogenetics identifies classes of specimens, the clades, by their shared derived expressions, the synapomorphies, which are also the potential biomarkers for the classes that they delimit. Synapomorphies are determined through polarity assessment (ancestral vs. derived) of m/z or gene-expression values and parsimony analysis; this process also permits intra and interplatform comparability and produces higher concordance between platforms. Furthermore, major trends in the data are also interpreted from the graphical representation of the data as a tree diagram termed cladogram; it depicts directionality of change, identifies the transitional patterns from healthy to diseased, and can be developed into a predictive tool for early detection.

Keywords

Biomarkers; Cancer; Early detection; Evolution; Omics; Parsimony; Phylogenetics

1 Introduction

Evolution is the unifying theme of all biological disciplines, and all explanations of biological phenomena should be compatible with evolutionary principles. Medicine cannot be an exception. Yet, the vast majority of publications that incorporate recent advances in genomics and proteomics are devoid of evolutionary reasoning and analytical methods. However, only recently there are new calls for the need of evolution in medicine in order to provide explanations for drug resistance in HIV and bacterial strains, autoimmune and degenerative diseases, as well as cancer typing and treatment [1-3]. Cancer development, progression, and maintenance are all evolutionary processes; they mirror similar evolutionary processes at the cellular and population levels in that they all involve genetic modifications, selective pressure,

Correspondence: Dr. Mones Abu-Asab, Laboratory of Pathology, National Cancer Institute, NIH, Bldg. 10/Rm 2A33, Bethesda, MD 20892, USA, **E-mail:** mones@mail.nih.gov, **Fax:** +1-301-480-9197. * Additional corresponding author: Dr. Hakima Amri; E-mail: amrih@georgetown.edu.

The authors have declared a conflict of interest. They will seek US patent rights for their UNIPAL algorithm.

and clonal propagation [2,4,5]. Therefore, evolution-compatible methods of analysis have a potentially useful role in cancer studies and diagnosis as well [2,6,7].

Evolutionary medicine seeks to explain the nature of disease in light of evolutionary theory [8]. It views the physicalities of the human body as a result of millions of years of natural selection that present compromises between differentiation at all levels and vulnerabilities [9]. Invoking evolution to explain medical phenomena will contribute to our understanding of how evolution works in diseases and how to counter with the proper treatment.

One of the earliest studies of disease etiology by evolutionary criteria was that of Sarnat and Netsky [5]. They described as “phylogenetic diseases” some of the degenerative and metabolic diseases that occurred in the derived structures of the mammalian brain. However, Azzone [4] attributed many diseases, such as cancer and autoimmunity, to mutations and their sustenance by natural selection, two processes that are at the crux of the evolutionary course of action. Since natural selective pressure is the main force determining diversity of living organisms and their state of health and disease, the data produced by omics (genomics, metabolomics, and proteomics) have to be analyzed in an evolutionary compatible way in order to produce biologically meaningful interpretations [2]. The tool that can bridge the gap between the omics data avalanche and evolutionary medicine is phylogenetics [10-12]. Phylogenetics is an analytical paradigm based on the principles of evolution. It has been employed by biologists in many disciplines such as botany, microbiology, and zoology, to construct relationships in an evolutionary sense at all the levels of the systematic hierarchy and more recently the tree of life [13]. Applying phylogenetic analysis to the omics data creates a paradigm shift where the evolutionary meaning of the data is brought out and applied to produce natural class determination, biomarker recognition, and modeling of the evolutionary processes of disease development. Furthermore, phylogenetic analysis is the evolutionary path between the omics data and their application in various practical settings. As the flowchart of Fig. 1 shows, there are only a few steps leading from raw data to applications: evolutionary polarity assessment of data values, phylogenetic algorithmic analysis, and interpretation.

2 The omics need phylogenetics

As it has become more evident recently, solving many of the problems in biomedical research is not going to be by producing more data, but rather by new methods of analyzing the data [13]. Today's omics data producing machines are sophisticated, sensitive, and accurate, and in the absence of human errors, their output is reliable and reproducible [14]. However, the over-reliance on parametric statistics for data analysis has reduced the useful inferences of patterns within the data [15]. Inferring of biological processes from data patterns is the main goal of bioinformatics [16], and a superior analytical tool has to be multidimensional. It must be able to accurately reveal biological patterns, processes, and classes; possess high predictivity; seamless and dynamic; able to combine several large datasets from multiple sources; suitable for intra and interplatform comparability; produce higher interplatform concordance; and its results can be utilized for early detection of disease, diagnosis, prognosis, and assessment of treatment [2].

This review will demonstrate how phylogenetics appears to be the most suitable analytical tool to provide the multidimensionality we are seeking, and its ability to translate the omics into a clinical tool.

3 Choosing between phylogenetics and phenetic clustering

There are two main schools of analytical thought in the bioinformatics of omics: the phenetic and the phylogenetic. The two differ on the relationship between the data values and the classification [17]. The phenetic school is very predominant in the analysis of microarray

expression data where it utilizes a distance matrix to produce dimensionless levels of difference [18]; the grouping of specimens is solely based on overall raw similarity without any evolutionary connotation (for an elaborate comparison of phenetic methods as applied to microarrays see Planet *et al.* [19]). On the other hand, the phylogenetic school attempts to reconstruct the biological processes from the known patterns within the data in an evolutionary meaningful sense. It sorts out the data into derived and ancestral states and uses only the derived ones to group specimens into a hierarchical model [10,11]. Because the two approaches differ conceptually and algorithmically, they seldom produce total congruency in their results (Fig. 2).

Phylogenetics appears to be very suitable for studying diseases with fast arising mutations, such as cancer, because it can modulate very recent divergence from normal conditions [20]. A phylogenetic analysis produces a hierarchical hypothesis of relationships (*i.e.* classification) among specimens that aims to reflect relatedness based on shared mutations and altered pathways. It reveals novel states of gene and protein expressions (these can be potential biomarkers, see below) and utilizes their distribution patterns among specimens for modeling their relatedness (*i.e.* groups specimens on the basis of their shared derived expressions but not overall similarity). Furthermore, it elucidates the direction of change among specimens that leads to their molecular and cellular diversity. The latter point is better illustrated with a graphical tree termed cladogram where the specimens with the highest number of novel expressions are located on the upper part of the tree (Fig. 2A).

To illustrate many of the fundamental differences between the two schools, we will examine one dataset analyzed by both methods. Figure 2 shows a phylogenetic cladogram (A) and a phenetic dendrogram (B), both are based on the same dataset [21]. To understand why the two trees have different explanations for the same data, we need to discuss the theoretical difference that they represent. Whereas the cladogram's hierarchy reflects the similarity between the specimens as based only on their shared derived expression values—data-based, the dendrogram uses Pearson's similarity coefficient of raw data (both ancestral and derived expression values)—specimen-based. Pearson's measures the correlation, r , between the specimens and produces a matrix of pair wise similarity ratios between the specimens; the average similarities are then calculated between groups of specimens to plot the dendrogram. Each node of the cladogram is based on the derived expression values that are shared by the specimens located above the node, and the segments' lengths bear no evolutionary significance (unless a molecular clock is assumed as in genomic-distance data). However, the relative lengths of dendrogram's segments are indicative of the percentages of shared average similarity between the specimens or groups of specimens.

Furthermore, among the significant differences between the two schools is that a phylogenetic analysis lessens the adverse effect of homoplasy while the phenetic does not [2,17]. Homoplasy is similarity due to convergence, parallelism, and reversal—all are evolutionary phenomena. Convergence occurs when two or more specimens have different developmental pathways for a homologous character state; parallelism is independently acquiring similar non-homologous states; and reversal is reverting to an ancestral state from a derived state. Homoplasies have a more detrimental effect on the phenetic analysis than the phylogenetic because they get an equal weight like all other similarity in the phenetic, while in the phylogenetic they compete against all other hypotheses of character distributions to generate the most plausible explanation of the data [22].

When considering information content, transmission, and retrieval, a phylogenetic classification is the most effective and efficient; it allows the storage of data in the smallest size diagram [17]. For example, because of its hierarchical nature, the cladogram indicates the direction of change from low to high, the dendrogram does not. This characteristic of a

cladogram gives significance to a specimen's position on the cladogram and the arrangements of branches. As we are going to see below, this allows the extrapolation from the cladogram to a clinical setting.

4 The omics meet phylogenetics

In the field of phylogenetics, a plethora of publications has accumulated during the last 30 years where phylogeneticists differed as to which approach is the optimum one for data analysis. There are three widely used methods to carry out phylogenetic analysis on omics data; these are Bayesian, likelihood, and parsimony. The first two methods are statistically based and related in that they require an explicit model of evolution such as homogenous rate of mutations, while parsimony is a non-statistical method that uses the minimum number of steps to explain the data. The three methods claim to produce the best hypothesis of relationships among a group of specimens [15,19]. Maximum likelihood method was devised by Felsenstein [23] to calculate the maximum likelihood function of a tree by incorporating specific assumptions (such as Markovian evolution and Poisson substitution) and branch lengths—times and mutation rates combined. The Bayesian approaches are derived from the likelihood method to measure the maximum posterior probability of individual trees by a sampling mechanism that incorporates branch lengths, substitution models, and their prior distribution [15].

When it concerns omics data, the choice of the analytical method to carry out a phylogenetic analysis is based on the optimum hypothesis of character states' distribution with highest fidelity to the data matrix, and obtaining the sought after information namely biomarkers, altered expressions and pathways, as well as disease classes (the clades). From a practical point of view, there is an obvious conflict between our stated goals and the first two statistical-based methods; their definitions of clades are irrelevant to medical interpretations, and their trees do not allow the tracing back of derived states, which are the potential biomarkers. Furthermore, the Bayesian and maximum likelihood approaches may be inefficient in dealing with heterogeneous rates of mutations and large number of specimens [15,24], could mistakenly attribute high probability to ambiguous groups, and may erroneously separate true sister groups because of their unequally long branches [25].

Maximum parsimony requires fewer parameters estimation than maximum likelihood [26], and functions better than Bayesian and likelihood when data are heterogeneous (*i.e.* have various rates of mutation such as in cancer) [24,27,28]. Currently, we lack any predictive model for most of the current diseases studied by omics to fulfill the parameters needed for a Bayesian or likelihood analysis. The question of whether some diseases like cancer follow a developmental model is still unanswered, although it is assumed that the specimens of a disease share common pathway aberrations. A recent phylogenetic analysis of MS proteomes of three cancers (ovarian, pancreatic, and prostate) has shown the probability of cancerous developmental models that transcend the three types [2]. Furthermore, data analysis of genomic and proteomic developments over a few clonal generations where the rate of change is fast and heterogeneous, is more suitable for parsimony analysis rather than Bayesian or likelihood.

By identifying uniquely shared derived states (the synapomorphies), parsimony analysis uncovers all the potential biomarkers within the dataset. It also defines natural classes, the clades, which are circumscribed by the synapomorphies (Fig. 2A). The most parsimonious tree it produces, the cladogram, maintains the same data pattern of the data matrix, therefore, the fidelity of the cladogram's character distribution to the original data is verifiable and can be extended to a clinical setting.

The state of currently used analytical tools for omics data places the scientists between two extremes, phenetic clustering and Bayesian/likelihood phylogenetics, without getting the expected rewards from either. Neither method seems to be the suitable paradigm for the omics

era. However, maximum parsimony appears to be the most optimum paradigm for the phylogenetic analysis of the omics data [15,28].

5 Predictivity of a parsimony phylogenetic analysis

A major problem that has characterized many omics-based studies is the provisional nature of their conclusions [29], also known in phylogenetics as low predictivity or lack of it. For a set of specimens, there is only one correct hypothesis of relationships that is based on their profiles; its predictive power is directly correlated with the robustness of the hypothesis—its capacity to fulfill its predictions. It is well established that a phylogenetic classification has a higher predictivity than a phenetic one [17]. Predictivity here is not the same as class prediction [30]. While class prediction deals with assigning a specimen to a known class, predictivity is the ability to list or predict a specimen's characteristics when its class becomes known (*i.e.* when biomarkers or an algorithmic analysis assigns the specimen to a class on the basis of its omics profile). Therefore, predictivity is a statement of accuracy on the hypothesis of relationships and its class definitions.

Predictivity is important when the hypothesis of relationships will be extended to and implemented in a clinical setting, *i.e.* translating the phylogenetic classification into practice. For example, high predictivity is needed in cancer diagnosis and prognosis where the classification of cancers has been mostly based on microscopy and a few immunohistochemistry markers. Furthermore, tumors with similar histopathology have shown divergent clinical courses and outcomes [6,30]. Applying evolutionary parsimony phylogenetics to cancer omics will produce a cancer classification that encompasses all types of available data and is expected to have the highest degree of predictivity. Therefore, having a predictive system of cancer classification brings higher objectivity to diagnosis and prognosis, as well as robust biomarker identification.

6 Shared derived patterns: synapomorphies

6.1 General remarks

Parsimony phylogenetics is based on the principle that shared derived patterns, the synapomorphies, can circumscribe natural groups called clades (*s. clade*). The shared derived patterns in the omics context may constitute a number of novel changes that occur in the specimens under study. These encompass all genetic mutations, novel and lost proteins, up- and down-regulated proteins, over and under-expressed RNA, as well as dichotomously asynchronous (DA) expression patterns of proteins and genes (see Section 8). A clade's members share one or more of these synapomorphies. For example, if only the specimens of pancreatic cancer share a unique mutation that is not shared by any normal specimens, then this mutation constitutes a synapomorphy, and specimens carrying the mutation are members of a clade. However, as explained below, there is an exact procedure for determining what constitutes a synapomorphy.

6.2 Evolutionary polarity assessment: identifying synapomorphies

Major omics techniques, such as MS proteomics and microarray, produce data in absolute values that impose limitations on their use and interpretation due mainly to inconsistency in reproducibility [14,29]. These data are usually utilized as similarity matrices for cluster analyses or probed with custom algorithms in search of novel values. There are major drawbacks for the direct use of absolute values in an analysis; prominent among them is the limitation on comparability within and between platforms as well as the lack of directionality within this form of data. Even the conversion of the data to similarity matrix by *t* and *F* statistics as well as fold-change results in significant loss of meaningful information [19,31].

The absolute data can be transformed into discrete values with evolutionary polarity assessment (EPA) [2]. In two-state and multistate characters, EPA is used to determine the proper evolutionary sequence of states, and consistent with this purpose, it is used here to sort out an absolute value into one of two states: ancestral or derived. Putting this in a mathematical context, the derived is given the value of 1, and 0 for the ancestral. The EPA process transforms the initial data into discrete binary states of 0s and 1s.

To implement EPA on a dataset, the experimental design should include a control subset of specimens; for example, when studying a cancer type, the control specimens should be healthy non-cancerous specimens. The control specimens will be used as the outgroup against which the values of the experimental specimens (the ingroup) will be compared. Table 1 illustrates how the process of polarity assessment is carried out for every m/z or gene-expression values of the experimental specimens. First, for every m/z point, the minimum and maximum values of the controls are determined—the range; if the value of an experimental specimen falls within the controls' range then it is considered ancestral and is assigned the value 0; if it falls outside the range, it is said to be derived and assigned a value of 1. Thus, the new transformed matrix is a polarized matrix with 0s and 1s.

It is clear here that the number of control specimens used for an outgroup comparison is an important criterion to correctly polarize the data and eliminate noise. For an analysis to be meaningful and provide high predictivity, the number of normal specimens that incorporates the maximum variation per population should be established [32]. An added advantage to the EPA data-transforming process is that it diminishes data inconsistency—a difficult to control noise that stems from several, mostly uncontrollable, factors during the experiment and data collection. However, noise reduction by EPA is handled by using control specimens in the experiment as the outgroup for polarity assessment.

There are theoretical and practical implications to transforming the data through EPA. For every specimen, the 1s represent the novel change that does not exist in the control outgroup, and therefore, may be indicative of a genetic mutation or protein modification depending on the data at hand. The 1s are called apomorphies (s. apomorphy); and if, all the experimental specimens have 1 for the same data point, then this data point is a shared derived state and is termed a synapomorphy. Therefore, all synapomorphies are potential biomarkers (see Section 7).

The 0s and 1s of a specimen make up its profile of ancestral and derived states. This profile determines the relatedness of the specimen to other specimens through the apomorphies they share—the synapomorphies. Therefore, class membership is determined by the competing number of synapomorphies among the specimens on the basis of maximum parsimony.

There are several new ways of utilizing the polarized data in analysis that are not attainable with the original absolute data such as pooling of datasets as well as intra- and interplatform comparability. Polarized data from one experiment can be directly subjected to an algorithmic analysis, or several polarized data from separate experiments with different specimens can be pooled together to produce an inclusive analysis. Furthermore, genomic, metabolomic, and proteomic data for the same set of specimens can be polarized separately, pooled in one matrix, and analyzed together to produce an inclusive analysis based on the three sets. Table 1 shows an example of pooling with two small real datasets. The number of polarized datasets that can be pooled together is theoretically limitless. This type of pooling is possible because polarized datasets have equal weight since each identifies the apomorphies of its specimens by discrete rather than absolute values.

The hypothetical example of Table 1 illustrates the pooling of two datasets, MS proteome and microarray for the same group of specimens. Each set was polarized using its own respective

set of controls, and then the two polarized sets (right side of A and B) were pooled and analyzed by a parsimony program. Although each of the two polarized datasets produced multiple equally parsimonious cladograms, the inclusive matrix produced one most parsimonious cladogram (C).

Using the data ranges of the control specimens provides higher stringency than using the statistical means (see also Section 8 on complex patterns below). By utilizing the controls' range as an ancestral criterion, every data value is evaluated individually to determine its evolutionary polarity; the ones falling within the controls' range are assigned an ancestral status. Using the statistical means of the experimental specimens by averaging their values does not exclude the values that fall within the controls' range, distorts the significance of their distribution, and prevents their tracking in the analysis. Furthermore, statistical means misrepresent data points that violate normal distribution—the ones with DA distribution (Fig. 3).

7 Evolutionary definition of biomarkers as synapomorphies

As biomarker discovery is a highly sought after criterion in the omics data, one will favor the analytical method that makes this process accurate, meaningful, and achievable. Parsimony phylogenetic analysis differs from likelihood and phenetic methods in that it maintains the identity of data points through the computational process, thus it allows the identification of every significant shared derived value that defines the natural groups of the hierarchical classification—the clades.

To carry out a parsimony phylogenetic analysis on a set of data, EPA of the data points is needed to determine whether a gene expression or m/z value is derived or ancestral. A shared derived state, a synapomorphy, among a number of diseased specimens is a potential biomarker for the group. Equating biomarkers with synapomorphies has an evolutionary connotation because it defines a natural group of specimens sharing similar expression and declares their ontogenic relatedness. This logical definition of what constitutes a biomarker requires a clear declaration that the biomarker is derived in relation to the controls and shared by all the members of its clade.

A synapomorphic biomarker can be supported by other synapomorphies that circumscribe the same clade; the higher the number of synapomorphies the higher the confidence in the predictivity of the selected biomarker. In addition, the occurrence of several synapomorphies for a clade offers more choices for selecting the optimum biomarker.

8 Incorporating complex patterns of omics

Proteomic and genomic data contain expression patterns that cannot simply be reduced to a statistical abstraction for data analysis. Such patterns are misrepresented when transformed into means, compared in fold-changes, or excluded from the analysis due to their complexity. One such pattern that is pervasive in cancer specimens is the DA distribution of gene expressions and proteins in a group of specimens [33]. The term dichotomous refers to a two-peak distribution with one peak above and the other below the range of control/normal specimens, while asynchronous denotes deviation from the normal range (Fig. 3). DA seems to be a population phenomenon that is only noticeable when a good number of specimens are included in the study.

Statistical methods of analysis usually average the values of the specimens in the study in order to carry out comparisons by either t- and F-statistics or fold-change, and therefore, misrepresent and overlook any meaningful interpretation of the distribution pattern of the DA expressions. The presence of several to many DA gene expressions and proteins in a set of specimens

underscores a complex pattern of pathway diversity that is difficult to model or classify in a simple phenetic clustering, but can be dealt with effectively and meaningfully in a parsimony phylogenetic context. Evolutionary polarity assessment for this phenomenon should consider it a multistate character and assigns different symbols for the values above and below the normals' range; a parsimony phylogenetic analysis will then deal with each of these states as independent from one another. For a discussion and examples of multistate coding and analysis see Felsenstein's instructions for using PHYLIP [34].

Although the DA pattern is a known phenomenon to scientists, and has been reported in tissues and cell lines [35,36], there has been no meaningful explanation for such variation or analytical considerations, and its evolutionary implications are still unknown. However, it may be related to the presence of several developmental pathways of cancer and other diseases [2].

9 Parsimony phylogenetic analysis of omics: an example

Currently available computing power permits the processing of large size matrices with relative speed. It is possible now to run a large data matrix with hundreds of specimens and tens of thousands of data points per specimen within a reasonable time [15]. During our experimentation with polarized proteomic data matrices on the parsimony program MIX, we managed to run a 23×10^6 -point matrix (180 specimens) in 18 h on a 3.2 GHz CPU.

We have selected the parsimony program MIX of Felsenstein [34] to carry out a maximum parsimony phylogenetic analysis because of its speed, reliability, available settings, and output format (MIX is freely available from <http://evolution.gs.washington.edu/phytip.html>). MIX is part of the PHYLIP package that contains a number of other applications that can be utilized for a number of phylogenetic analyses for the same dataset such as likelihood and distance. All of these programs are controlled by a menu that allows options for the analysis.

There are only a few examples of phylogenetic analysis of omics data [2,6,7,37]. One of the practical problems for many researchers that limit their usage of phylogenetic programs is the transformation process of the raw data to an input format that is acceptable by the program. Figure 4 presents an example of a parsimony phylogenetic analysis of MS SELDI-TOF proteomic serum data of prostate cancer patients and healthy men. The normalized raw data was polarized according to the polarity assessment method described in Section 6.2 and explained by an example in Table 1. However, this process was automated and carried out here by a computer program written by the authors (UNIPAL, Universal Polarity Assessment Algorithm) [2].

UNIPAL transformed the original raw data into a matrix of 0s and 1s for all the cancerous specimens by using the normal specimens' range for every m/z point as the baseline for polarity assessment (outgroup comparison). Then, the newly produced polarized matrix was processed with MIX to run a Wagner maximum parsimony. There are two output files produced by MIX, both in text format and can be read by any text reader program: the Outfile and the Treefile. The Outfile has all the equally parsimonious trees that data supports in a graphical format; and a list of all the synapomorphies supporting every node can be produced if one invokes this option. The Treefile lists the same trees as in the Outfile in a text format; this can be used to draw and edit the trees in other programs such as TreeView [38].

MIX produced one most parsimonious cladogram for this dataset that is based on 36 prostate cancer patients and 49 healthy men (Fig. 4). One most parsimonious cladogram means that MIX found only one tree that has the smallest number of steps to explain the relationship between the specimens. With other datasets, MIX may produce several equally parsimonious cladograms with some variation in minor branches. An interpretation of the topology of the cladogram and what it reveals about the trends in the dataset is discussed below.

10 The structure of an omics cladogram

The cladogram is the graphical representation of the hypothesized hierarchical relationships among the specimens that defines classes of specimens. The tip of each line of the cladogram denotes a specimen. It is the most efficient summary of the information contained in the raw data [17]. Each node on the cladogram is justified by the shared derived state(s) among the specimens of one of the segments.

The topology of the cladogram also conveys general trends within the data that are not obvious otherwise by other types of analysis [2]. Our own analyses of several MS proteomic and microarray data have indicated that there are three distinct sections of the cladogram: the basal, the middle, and upper. The basal contains most of the normal specimens; the middle has the transitional specimens between the normal and cancer, and the upper section has the cancerous ones. Figure 4 shows a parsimonious cladogram of MS proteomic serum specimens that were taken from healthy and diseased men. Its upper section has a dichotomy defining two major clades of the prostate cancer; both are more or less equal in size. The basal section is restricted to healthy specimens; it has a well-defined large clade encompassing the majority of specimens, one minor clade below the large one, and a few single-specimen clades at the bottom. Additionally, the middle section has mostly single-specimen clades in tandem that are of normal specimens and cancerous ones. The lower part of this section has the normal clades and upper part has the cancerous clades.

Similar cladogram topology exists so far in all large genomic and proteomic datasets that we have analyzed thus far [2]. This makes the cladogram the only tool thus far that could identify the transitional patterns from healthy to cancerous, and possibly renders it a predictive tool for early disease detection.

11 Testing the congruence of omics' data

Incongruity of omics data is a criterion that has become a topic of serious debate [14,29], and the field is in need of robust method for testing congruence. The parsimony phylogenetic analysis as described here offers an evolutionary approach for testing concordance of datasets. In addition to carrying out inclusive analysis by the pooling of multiple omics datasets, several other data processes are possible under this model. Thus far we have focused in our presentation and discussion on examples of high-throughput data experiments, the approach outlined here is also applicable to many other types of data such as 2-D gels, as well as chromosomal and genomics data. As long as data polarity can be determined, a parsimony analysis is most likely achievable.

Interplatform comparability is attainable here at two levels: first, by testing the congruence of the data from two or more sources for the same set of specimens (for example, does proteomic data produce the same classification as genomic data?); secondly, by testing the congruence of the synapomorphies (the potential biomarkers) among different sets of specimens. Having multiple datasets for the same specimens allows the testing for the data accuracy and the robustness of the classification hypothesis [20,39].

Small sample size and variable methods of analysis are two chronic problems in published studies and experimental designs, which prevent direct comparisons of the results and conclusions. However, applying EPA and parsimony analysis will enable us to test the congruity of an experiment at the two levels mentioned above. The phylogenetic model produced higher congruence in synapomorphies of two separate sets of specimens from the same tissue type. We tested the concordance of two published studies [21,40] on uterine fibroids (leiomyomas) by comparing their two lists of synapomorphies, and found 62% concordance in synapomorphies despite the variation in the number of probes between the two

datasets; which was greatly enhanced in comparison with the 13% concordance shown between the published statistical gene lists of the two studies.

Because of its hierarchical nature, phylogenetic classification makes it possible to test gene linkage and specific alterations to pathways. The path of synapomorphies from the base of the cladogram to its tip is a sequential developmental map for the successional events that produce the different stages of the disease and the diversity of its specimens. These synapomorphies are the shared derived alterations that we need to identify since they will elucidate the disease etiology and are the biomarkers of its various stages.

12 Translating phylogenetic analysis of omics into clinical practice

The practical aspect of phylogenetic analysis can be realized in various ways: better diagnosis of diseases and disorders through evolutionary classification that reflects the real ontogeny and phylogeny of disease, better treatment by fine targeting of pathways, and better assessment of health status from faster and cost-effective omics data analysis. The following illustrates through theoretical clinical scenarios the potential applications of parsimony phylogenetic analysis of omics in a clinical setting.

Scenario A: Routine health assessment

The health status of the individual can be routinely assessed during a routine checkup from a small blood specimen (<0.5 mL) for early detection of degenerative diseases and cancer. The serum fraction is submitted for proteomic MS analysis, and the spectra are analyzed using parsimony phylogenetics against the serum of control specimens (healthy and diseased). The location of this individual on the cladogram (within healthy, diseased, or transitional clades, Fig. 5) determines the health status (healthy, diseased, or transitioning from healthy to disease). A specimen of a healthy individual assembles within the healthy clades.

Scenario B: Early detection and prevention

Individuals located within the transitional clade, nested between the healthy and cancer specimens (in this example, Fig. 5), are at-risk of developing cancer. Therefore, the at-risk individuals are accumulating mutations that are making them susceptible but have not yet reached clinical manifestations. In the evolutionary medical paradigm offered by the parsimony phylogenetic analysis, this person would be “at-risk” of developing disease/cancer. Preventive medicine could play a major role in this case.

Scenario C: Diagnosis

If the phylogenetic analysis places the individual's proteomic specimen within the cancer clades (Fig. 5), then the patient is a cancer carrier. We have demonstrated before, but have not yet validated, that for three cancer types (ovarian, pancreatic, and prostate) each produced its own clades separate from the other two, therefore, it is possible in a comprehensive analysis to place the patient within the respective type of cancer [2].

Scenario D: Post-treatment evaluation and prognosis

Depending on the position of the patient's proteomic specimen within the cancerous clades (basal, middle, or terminal) of the analysis cladogram, the cancer clinical stage can be determined. The staging here is based on the derived mutations that patient carries—the apomorphies.

After a course of treatment – chemotherapy, radiation, surgery – the patient's progress can be evaluated by a proteomic-phylogenetic analysis of their serum. If the location of the patient's specimen on the analysis cladogram has moved from the cancerous clades to the normal, then

the treatment has succeeded. Follow up and monitoring can be carried out periodically by this minimally invasive and cost-effective method.

13 Conclusions

Evolutionary analysis and interpretation of the omics data offer a unifying paradigm for the various types of the data and provides a multidimensional application of the analysis in a medical context. Cellular processes involved in disease development recapitulate evolutionary processes; they involve genetic modifications, selective pressure, and clonal propagation. Furthermore, diseases are currently assumed to be natural classes and subclasses with each having its own unique aberrations in developmental pathways. Thus, employing of evolutionary polarity assessment to sort out uniquely derived omics states coupled with parsimony phylogenetic analysis seem to provide a predictive, seamless, and dynamic evolutionary classification of the specimens that accurately reveal biological classes, patterns, and processes. This parsimonious paradigm is also capable of combining several large datasets from multiple sources for inclusive analyses, produces higher interplatform concordance, and offers intra and interplatform comparability. Additionally, a parsimonious cladogram reveals the directionality of change within a set of specimens, and could be utilized for early detection, diagnosis, prognosis, assessment of treatment, and biomarker identification. The parsimony phylogenetic approach could also serve as the basis for the individualized medicine of the 21st century.

References

1. Nesse RM, Stearns SC, Omenn GS. Medicine needs evolution. *Science* 2006;311:1071. [PubMed: 16497889]
2. Abu-Asab M, Chaouchi M, Amri H. Phyloproteomics: what phylogenetic analysis reveals about serum proteomics. *J Proteome Res* 2006;5:2236–2240. [PubMed: 16944935]
3. Shackney SE, Silverman JF. Molecular evolutionary patterns in breast cancer. *Adv Anat Pathol* 2003;10:278–290. [PubMed: 12973049]
4. Azzone GF. The nature of diseases: evolutionary, thermodynamic and historical aspects. *Hist Philos Life Sci* 1996;18:83–106. [PubMed: 8940904]
5. Sarnat HB, Netsky MG. Hypothesis: Phylogenetic diseases of the nervous system. *Can J Neurol Sci* 1984;11:29–33. [PubMed: 6704791]
6. Pennington G, Smith CA, Shackney S, Schwartz R. Expectation-maximization method for reconstructing tumor phylogenies from single-cell data. *Comput Syst Bioinformatics Conf* 2006:371–380. [PubMed: 17369656]
7. Desper R, Khan J, Schaffer AA. Tumor classification using phylogenetic methods on expression data. *J Theor Biol* 2004;228:477–496. [PubMed: 15178197]
8. Nesse RM. How is Darwinian medicine useful? *West J Med* 2001;174:358–360. [PubMed: 11342524]
9. Culotta E, Pennisi E. Breakthrough of the year: evolution in action. *Science* 2005;310:1878–1879. [PubMed: 16373538]
10. Felsenstein, J. *Inferring phylogenies*. Sinauer Associates; Sunderland, MA: 2004.
11. Wiley, EO. *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. John Wiley and Sons; New York: 1981.
12. Hennig, W. *Phylogenetic systematics*. University of Illinois Press; Urbana: 1966.
13. Whitfield J. Linnaeus at 300: we are family. *Nature* 2007;446:247–249. [PubMed: 17361152]
14. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006;24:1151–1161. [PubMed: 16964229]
15. Goloboff, PA.; Pol, D. Parsimony, phylogeny, and genomics. Albert, VA., editor. Oxford University Press; Oxford, New York: 2005. p. 148–159.
16. Gascuel, O. *Mathematics of evolution and phylogeny*. Gascuel, O., editor. Oxford University Press; New York: 2005. p. 1–8.

17. Farris JS. The information content of the phylogenetic system. *Syst Zool* 1979;28:483–519.
18. Albert, VA. Parsimony, phylogeny, and genomics. Albert, VA., editor. Oxford University Press; Oxford, New York: 2005. p. 1–11.
19. Planet PJ, DeSalle R, Siddall M, Bael T, et al. Systematic analysis of DNA microarray data: ordering and interpreting patterns of gene expression. *Genome Res* 2001;11:1149–1155. [PubMed: 11435396]
20. Kumazawa Y, Nishida M. Sequence evolution of mitochondrial tRNA genes and deep-branch animal phylogenetics. *J Mol Evol* 1993;37:380–398. [PubMed: 7508516]
21. Quade BJ, Wang TY, Sornberger K, Dal Cin P, et al. Molecular pathogenesis of uterine smooth muscle tumors from transcriptional profiling. *Genes Chromosomes Cancer* 2004;40:97–108. [PubMed: 15101043]
22. Mickevich MF. Taxonomic congruence. *Syst Zool* 1978;27:143–158.
23. Felsenstein J. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biol J Linnean Soc* 1981;16:183–196.
24. Stefankovic D, Vigoda E. Phylogeny of mixture models: robustness of maximum likelihood and non-identifiable distributions. *J Comput Biol* 2007;14:156–189. [PubMed: 17456014]
25. Siddall ME. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. *Cladistics* 1998;14:209–220.
26. Goloboff PA. Parsimony, likelihood, and simplicity. *Cladistics* 2003;19:91–103.
27. Stefankovic D, Vigoda E. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Syst Biol* 2007;56:113–124. [PubMed: 17366141]
28. Kolaczowski B, Thornton JW. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 2004;431:980–984. [PubMed: 15496922]
29. Coombes KR, Morris JS, Hu J, Edmonson SR, Baggerly KA. Serum proteomics profiling—a young technology begins to mature. *Nat Biotechnol* 2005;23:291–292. [PubMed: 15765078]
30. Golub TR, Slonim DK, Tamayo P, Huard C, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531–537. [PubMed: 10521349]
31. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 2006;7:55–65. [PubMed: 16369572]
32. Graybeal A. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol* 1998;47:9–17. [PubMed: 12064243]
33. Lyons-Weiler J, Patel S, Becich MJ, Godfrey TE. Tests for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics* 2004;5:110. [PubMed: 15307894]
34. Felsenstein J. PHYLIP: Phylogeny Inference Package (Version 3.2). *Cladistics* 1989;164–166.
35. Fulda S, Poremba C, Berwanger B, Hacker S, et al. Loss of caspase-8 expression does not correlate with MYCN amplification, aggressive disease, or prognosis in neuroblastoma. *Cancer Res* 2006;66:10016–10023. [PubMed: 17047064]
36. Reed JC, Meister L, Tanaka S, Cuddy M, et al. Differential expression of bcl2 protooncogene in neuroblastoma and other human tumor cell lines of neural origin. *Cancer Res* 1991;51:6529–6538. [PubMed: 1742726]
37. Uddin M, Wildman DE, Liu G, Xu W, et al. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proc Natl Acad Sci USA* 2004;101:2957–2962. [PubMed: 14976249]
38. Page RD. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 1996;12:357–358. [PubMed: 8902363]
39. Miyamoto MM, Fitch WM. Testing species phylogenies and phylogenetic methods with congruence. *Syst Biol* 1995;44:64–76.
40. Hoffman PJ, Milliken DB, Gregg LC, Davis RR, Gregg JP. Molecular characterization of uterine fibroids and its implication for underlying mechanisms of pathogenesis. *Fertil Steril* 2004;82:639–649. [PubMed: 15374708]

Abbreviations

DA	dichotomously asynchronous
EPA	evolutionary polarity assessment

Glossary

Definitions of terms

Clade

a group of specimens sharing one or more synapomorphies

Cladogram

a graphic classification of hierarchical relationships among specimens based on the synapomorphies (shared derived characters). It is also a summary of trends that occur within the data, and shows the directionality of accumulation of change with the highest number of synapomorphies shared by the specimens that are closer to the upper part of the cladogram

Dendrogram

a tree diagram used to graphically illustrate the arrangement of the clusters produced by a phenetic clustering algorithm (see Phenetic Clustering). Dendrogram is produced in computational biology (*e.g.* microarray analysis) to illustrate similarity and gene-linkage

Dichotomous asynchronicity

a two-tailed pattern of protein or gene expression in a number of specimens with a physiological abnormality (*e.g.* cancer) in comparison with the normal specimens. Usually, the values of *m/z* protein or gene expression of the abnormal specimens are outside the range of the normal specimens (*i.e.* above and below the normal range)

Dynamic classification

a classification that has the capacity to incorporate novel specimens without major alterations to the composition of its main groups or their relationships

Evolutionary medicine

a branch of medicine that seeks to explain the nature of disease in an evolutionary context

Homoplasy

similarity due to convergence, parallelism, or reversal. Convergence occurs when two or more specimens have different developmental pathways for a homologous character state; parallelism is independently acquiring similar non-homologous states; and reversal is reverting to an ancestral state from a derived state

Ingroup

the group of specimens under study, for example, diseased specimens

Interplatform comparability

the ability to compare several datasets produced on different platforms. Evolutionary polarity assessment transforms the omics data into polarized matrices of discrete values (0/1), therefore, makes possible the ability to compare

two or more separate experiments, and/or to combine several experiments in one large analysis

Interplatform concordance

the intersection (sharing) of significant m/z values and gene-lists produced by two or more separate experiments. The higher the concordance between independent experiments the larger the number of common proteins and genes, and the more significant the results

Mass to charge ratio (m/z)

a unique value of a protein's mass (m) to the total charge (z) it carries. The m/z is the value obtained from laser desorption mass spectrometry machines such as SELDI or MALDI

Outgroup

a group of specimens used to polarize the ingroup values of m/z or gene expression into ancestral (plesiomorphic) and derived (apomorphic)

Parsimony

means simplicity, the preferred hypothesis is the one requiring the least number of explanations (Occam's Razor). In the omics context, the preferred phylogenetic cladogram is the one that requires the least number of steps to construct it from the polarized data matrix

Phenetic clustering

grouping specimens on the basis of similarity without a priori sorting of similarity into ancestral and derived. Phenetics does not provide any information about the evolutionary phylogenetic relationships among specimens

Phylogenetic classification

a classification that uses synapomorphies to delimit clades (*i.e.* monophyletic groups), it provides evolutionary phylogenetic relationships among specimens

Polarity assessment

also known as outgroup comparison. It is the basis of sorting out the data values (whether proteomic [m/z], or microarray expression values) into ancestral and derived. In large datasets, it transforms absolute numbers of data values into polarized binary numbers (0/1), where zero (0) signifies ancestral and one (1) signifies derived

Predictive classification

a classification that reveals the characteristics of a specimen when its place in the classification is determined

Synapomorphy & biomarker

a shared derived protein or gene expression value in comparison with a number of normal specimens (the outgroup). A protein synapomorphy may have one of the following conditions: (i) a novel protein, (ii) a disappeared protein, (iii) up-regulated protein, (iv) down-regulated protein, and (v) dichotomously asynchronous regulated protein (the m/z values are above and below the normals' range but not within the normals' range). A gene synapomorphy may have one of the following conditions: (i) over-expressed value above normals' range, (ii) under-expressed value below the normals' range, (iii) dichotomously asynchronous values, and (iv) undetectable expression value. Biomarkers are unique proteins or gene expressions that could delimit (or characterize) a group of specimens sharing a physiological condition. Because synapomorphies also

group together specimens sharing uniquely derived protein or gene expression into clades (*i.e.* every clade has its own set of synapomorphies), these synapomorphies are potential biomarkers for the clade they define

Figure 1.

Flowchart outlining the various stages of an evolutionary phylogenetic analysis of omics data, as well as the interpretation and translation of the analysis results into a clinical setting.

Figure 2.

Phylogenetics *vs.* Phenetics. (A) Phylogenetic cladogram based on maximum parsimony analysis, and (B) phenetic dendrogram based on Pearson's correlation for the same data set [21]. While the cladogram resolves the relationship between the leiomyoma and leiomyosarcoma specimens by finding 32 uniquely expressed synapomorphies shared by both groups and 20 synapomorphies distinguishing leiomyosarcomas from the leiomyomas, the dendrogram fails to resolve this relationship and clusters the leiomyomas with normal myometrium specimens. The cladogram has directionality for accumulated synapomorphies, and the dendrogram does not. For example, the cladogram indicates that the leiomyosarcoma specimen GSM11779 has the highest number of synapomorphies, and GSM11769 has the lowest. Dataset GDS533 available at <http://www.ncbi.nlm.nih.gov/geo/>.

Figure 3. Dichotomously asynchronous protein and gene-expression. Two-tailed distribution that occurs in a group of cancerous specimens. (A) Protein intensity at m/z 12 215 of 11 specimens of prostate cancer and 17 normals; six cancerous specimens show upregulation and five downregulations. (B) RNA signal intensity of ten specimens of uterine leiomyosarcoma showing four specimens overexpressing and six underexpressing Akt1.

Figure 4.

A most parsimonious cladogram produced by MIX for MS serum proteomic data of 36 prostate cancer patients and 49 healthy men. Each specimen had 15 144 m/z data points; polarity assessment was carried out by UNIPAL. Each line that ends on the right side of the figure represents a specimen. The red part of the cladogram indicates the cancerous specimens as diagnosed before the experiment; the green section indicates the healthy specimens; and the blue shows the presumed healthy specimens that seem to form a transitional zone between the healthy and cancerous clades.

Figure 5.

Translating the parsimony phylogenetic analysis of omics into clinical practice. A schematic topology of a typical proteomic cladogram of a cancer analysis. There are two major cancerous clades at the upper section of the cladogram; transitional clades in the middle section; and the basal healthy clades. Adding an unknown specimen to an analysis will have three possible scenarios: scenario A indicates the likely location of a healthy specimen within the healthy clades; scenario B places a specimen from a susceptible individual with the transitional clades between the healthy and cancerous clades; and scenario C would locate a cancerous specimen within one of the two major cancer clades. A post-treatment analysis may change the location from the cancerous to transitional or healthy clades depending on the treatment's efficacy.

Table 1

From omics to cladogram. The process of evolutionary polarity assessment is illustrated by using a sample of MS proteomic data, (A), and gene expression data, (B). Each dataset consists of ten specimens: five controls and five experimental. The minimum and maximum (*i.e.* range) of the controls are determined for each *m/z* point and gene, and then intensity values of the experimental specimens are transformed to either 0 or 1 depending on whether its value is within the controls' range, or outside, respectively. The polarized values from the proteomic data, (A), and gene-expression, (B), are pooled together and processed in an algorithmic parsimony analysis (MIX) to produce the consensus cladogram (C). The synapomorphies for each clade are listed at its node

A. MS Proteomic Data																	
m/z	Controls							Intensity							Polarized Values of Experimentals		
	Controls							Experimentals									
	A	B	C	D	E	Min	Max	F	G	H	I	J	F	G	H	I	J
928.70	115	59	163	137	131	59	163	45	62	62	577	145	1	0	0	1	0
979.43	222	145	273	263	256	145	273	97	111	111	379	238	1	1	1	1	0
1018.58	145	76	225	200	196	76	225	72	101	75	665	208	1	0	1	1	0
1110.48	200	126	257	271	186	126	271	81	85	102	319	225	1	1	1	1	0
1228.72	105	84	223	180	176	84	223	69	91	72	222	202	1	0	1	0	0
1304.16	142	94	211	201	134	94	211	79	90	66	123	186	1	1	1	1	0
1407.12	114	70	217	196	150	70	217	62	96	72	109	204	1	0	0	0	0
1511.54	143	111	249	241	211	111	249	75	120	85	157	258	1	0	1	0	1
1623.71	245	94	263	214	183	94	263	73	115	95	151	200	1	0	0	0	0
1711.73	157	91	222	183	164	91	222	71	121	72	115	179	1	0	1	0	0

B. Gene Expression Data																	
Gene	Controls							Experimentals							Polarized Values of Experimentals		
	Controls							Experimentals									
	A	B	C	D	E	Min	Max	F	G	H	I	J	F	G	H	I	J
PTAFR	18	70.1	21.2	11.6	19.7	11.6	70.1	93.4	29	9.1	34.4	89.6	1	0	1	0	1
PRKCD	371.6	418.8	375.9	447.5	353	353	447.5	239.4	178.8	260.1	267	270.2	1	1	1	1	1
ACAT1	184.6	189	200.2	217	235.1	184.6	235.1	196.8	144.4	321.2	101.6	138.7	0	1	1	1	1
MARCKS	446.4	458.2	493.2	556.5	372.1	372.1	556.5	324.3	340.7	338.8	437.2	501.2	1	1	1	0	0
OGDH	142.3	145.3	238.5	187.5	162.8	142.3	238.5	171.9	152.2	123.4	158.4	175.8	0	0	1	0	0
CRK	22.8	28	28.7	32.1	25	22.8	32.1	41.6	43.4	11.1	28.5	23.8	1	1	1	0	0
CHKA	175	140.1	110.8	159.8	180.1	110.8	180.1	121.9	70.3	215.6	80	130.6	0	1	1	1	0
GPR109B	44.8	29	30.8	27.3	41	27.3	44.8	147.4	65.8	53.9	18.6	35.1	1	1	1	1	0
HTR1B	90.7	72.6	37.3	40.7	48.4	37.3	90.7	38.1	66.7	42.7	47.2	57.1	0	0	0	0	0
IL2RG	419.4	356.7	427.3	332.5	399.4	332.5	427.3	297.7	565.6	169.7	287.4	387.7	1	1	1	1	0

A. MS Proteomic Data																	
m/z	Intensity										Polarized Values of Experimentals						
	Controls					Experimentals											
	A	B	C	D	E	Min	Max	F	G	H	I	J	F	G	H	I	J
	P R K C D	G P R I 0 9 1 0 B	M A R C C R S K	1 7 2 1 2 8 1	1623 1407												

c.

The graph displays the following data points and trends:

- Peptides and their m/z values:**
 - M: 17
 - A: 21
 - R: 28
 - C: 100
 - K: 123
 - S: 1407
 - G: 1623
 - P: 191
 - R: 210
 - K: 290
 - C: 310
 - D: 330
- Peptides and their m/z values:**
 - F: 1623
 - H: OGDH
 - G: 1623
 - I: 1623
 - J: 1623