Biochimica et Biophysica Acta 1867 (2017) 101–108

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

PhyloOncology: Understanding cancer through phylogenetic analysis☆

Jason A. Somarelli [a,*], Kathryn E. Ware [a], Rumen Kostadinov [b], Jeffrey M. Robinson [c,d], Hakima Amri [e], Mones Abu-Asab [f], Nicolaas Fourie [d], Rui Diogo [c], David Swofford [g], Jeffrey P. Townsend [h,i,j,**]

[a] Duke Cancer Institute and the Department of Medicine, Duke University Medical Center, Durham, NC 27710, United States
[b] Pediatric Oncology, School of Medicine, Johns Hopkins University, United States
[c] Anatomy Department, College of Medicine, Howard University, Washington, DC 20059, United States
[d] Digestive Disorders Unit, National Institute of Nursing Research, NIH, Bethesda, MD 20892, United States
[e] Department of Biochemistry and Cellular and Molecular Biology, Georgetown University Medical Center, Washington, DC 20007, United States
[f] Section of Ultrastructural Biology, National Eye Institute, NIH, Bethesda, MD 20892, United States
[g] Department of Biology, Duke University Trinity College of Arts and Sciences, Durham, NC 27710, United States
[h] Department of Biostatistics, Yale University, United States
[i] Department of Ecology and Evolutionary Biology, Yale University, United States
[j] Department of Program in Computational Biology and Bioinformatics, Yale University, United States

ARTICLE INFO

Article history:
Received 20 July 2016
Received in revised form 14 October 2016
Accepted 26 October 2016
Available online 31 October 2016

Keywords:
Tumor heterogeneity
Cancer stratification
Clonal evolution
Tumor trees
Cancer types

ABSTRACT

Despite decades of research and an enormity of resultant data, cancer remains a significant public health problem. New tools and fresh perspectives are needed to obtain fundamental insights, to develop better prognostic and predictive tools, and to identify improved therapeutic interventions. With increasingly common genome-scale data, one suite of algorithms and concepts with potential to shed light on cancer biology is phylogenetics, a scientific discipline used in diverse fields. From grouping subsets of cancer samples to tracing subclonal evolution during cancer progression and metastasis, the use of phylogenetics is a powerful systems biology approach. Well-developed phylogenetic applications provide fast, robust approaches to analyze high-dimensional, heterogeneous cancer data sets. This article is part of a Special Issue entitled: Evolutionary principles - heterogeneity in cancer?, edited by Dr. Robert A. Gatenby.

1. Introduction

Cancer results from a breakdown in multicellular cooperation [1], evolving changes in DNA sequence, gene expression patterns, and/or epigenetic modifications that permit unchecked growth. These molecular changes induce phenotypes that can increase the ability of a cell to compete, survive and reproduce, and ultimately lead to cancer. Advantageous phenotypes include 1) self-sufficiency in growth signals, 2) insensitivity to anti-growth signals, 3) evasion of apoptosis, 4) limitless replicative potential, 5) sustained angiogenesis, 6) ability to invade and metastasize to surrounding tissue and distant organs, 7) deregulated cellular energetics, and 8) avoidance of immune destruction [2,3]. In many cases, these hallmarks are the consequences of mutations that result in a cell with increased fitness compared to its healthy counterparts, followed by selective pressures that increase the prevalence of that cell lineage. Continued rounds of mutation and selection putatively lead to more extreme phenotypes in comparison to normal tissue, and thereby more aggressive metastatic disease.

From the initial transforming event to dissemination, seeding, and eventual metastatic colonization, cancer progression represents a process of selection over time. Nowell first drew this parallel between the selective forces acting on cancer cells within the body and those acting on individuals within populations in nature [36]. Nowell proposed that the heterogeneity observed in tumors is due to an increase in genetic instability as cancer progresses [36]. Indeed, evidence of increased genetic instability over time has been recently shown in the progression of Barrett's esophagus to esophageal adenocarcinoma in a longitudinal study of patients for over 20 years [37]. This increased genetic instability enhances the genetic diversity of the cancer cell population, and presumably the phenotypic diversity as well, which is acted upon by selective forces within the tumor, such as immune surveillance, hypoxia, glucose deprivation, and the production of reactive oxygen species, to produce sub-clones capable of thriving despite the barriers to progression [38]. These concepts of increasing heterogeneity coupled with selection in the context of cancer progression have been borne out by studies using both first generation and next-generation sequencing technologies [39]. For example, analysis of breast cancer primary and

☆ This article is part of a Special Issue entitled: Evolutionary principles - heterogeneity in cancer?, edited by Dr. Robert A. Gatenby.

* Corresponding author.
** Correspondence to: J.P. Townsend, Department of Biostatistics, Yale University, United States.
E-mail addresses: jason.somarelli@duke.edu (J.A. Somarelli), Jeffrey.Townsend@Yale.edu (J.P. Townsend).

http://dx.doi.org/10.1016/j.bbcan.2016.10.006
0304-419X/

matched metastases found that just over half of the coding mutations identified in the metastases (19/32) were not detected in the primary tumor [40]. Of the mutations common to the primary tumor and metastases, 6/13 were found in only 1–13% of cells in the primary tumor [40]. Similarly, sequencing of pancreatic cancer primary tumors and metastases revealed that different metastases are seeded from unique clones [41]. These authors also concluded that metastatic clones may have seeded tertiary subclones [41], though the physical history of metastatic departures cannot be inferred from sequence data without a complete sampling of clonal lineages within the primary tumor [42]. In the largest multi-region sampling paper published to date, sampling of 40 patients with primary tumors and 3–8 matched metastases demonstrated diverse patterns of molecular genetic divergence along the time course of cancer progression [28]. These and other studies clearly demonstrate that cancer progression, from indolent neoplasia to aggressive and metastatic disease, is a process in which cells change in a spatio-temporal manner while under selective forces.

Given that cancer progression is governed by selective forces, tools developed to elucidate evolutionary relationships should generally be appropriate for use in the analysis of cancer. One of the most well-developed and successful evolutionary approaches is phylogenetics. Originally designed to model and infer evolutionary relationships among organisms, this suite of algorithms, concepts, and tools has been usefully applied in a wide array of diverse fields, even fields in which the data have no true evolutionary context [43–47]. Below, we briefly review phylogenetic concepts and methods and discuss the possibilities for the application of phylogenetics to analysis of cancer data sets in the following three capacities: 1) as a suite of classification algorithms that could be applied to assign specimens as coming from either healthy individuals, patients with localized disease, or those with metastasis; 2) as a means to deconstruct the complex heterogeneity within tumors; and 3) as a natural method to determine the branching evolution of cancer cells within individuals during cancer progression.

2. Phylogenetics: revealing relationships between states

The field of phylogenetic systematics was born from a need to sort and classify organisms in such a way as to capture their relationships by descent. Phylogenetics utilizes a data matrix of input characteristics from a group of organisms (Fig. 1A) to produce a graphical "tree" (Fig. 1B) where the branching pattern, or tree-topology, represents bifurcations between individuals, species, or higher taxa, depending on the scope of the taxonomic question of interest. A phylogeny, or an evolutionary tree, provides a basic structure to statistically analyze the evolutionary relationships (differences and similarities) among distantly-related taxa (species or larger groups of inclusively-related species). The most recent common ancestor (MRCA) of a group is the node furthest from the root that contains all members of the group as descendants. Pairs of taxa that share a more recent common ancestor are more closely related than those whose MRCA occurs more deeply in the tree (Fig. 1B). When numerous taxon divergences are represented along a lineage, it becomes possible to chart the accumulation of traits or features that have resulted from evolution over time. Tree topologies can be rooted or unrooted. In a rooted tree, some extrinsic information is used to root the tree. This information is typically in the form of an assumed outgroup. Outgroups typically represent distantly-related taxa that provide information on the ancestral condition (state) of a character prior to its transformation to a more derived condition. In principle, this use of outgroups enables the researcher to establish the directionality of change for a set of characters [48], though it must be cautioned that outgroups often have undergone significant evolutionary change themselves and are not always suitable proxies for the ancestor. Unlike rooted trees, unrooted trees reveal the relatedness of taxa within the nodes of the tree without assuming a relationship of a group of taxa to an ancestral state. Phylogenetic algorithms are widely used in the study of the evolutionary dynamics of molecular sequences themselves, using each homologous position in a sequence as an individual character [49–51]. Moreover, phylogenetic methods can be applied to numerous types of data, including morphological characteristics and/or other data that can be converted into discrete character states, and even can be applied to quantitative characters under appropriate models of evolutionary change.

3. Phylogenetic analysis of cancer data

The multiple and diverse paths of progression to cancer are forged by genetic mutations and alterations to epigenetics, gene expression, and protein signaling. Because these changes tend to accumulate over time in diversifying somatic lineages, phylogenetic analysis provides a natural tool set for evaluating the branching history of cancer onset and progression. The development of these tools has been considered an emerging field of inquiry, termed herein as *PhyloOncology* or *Cancer Phylogenetics*, which represents diverse applications of phylogenetic algorithms to the analysis of cancer data. We highlight below three general types of analyses in which phylogenetics has provided insight in the understanding of cancer biology: 1) classification of cancer specimens (Fig. 1C and D); 2) analyses of intratumoral heterogeneity (Fig. 1E and F), and 3) tracking clonal evolution and progression (Fig. 1E and F).

3.1. Applying phylogenetic methods to classify gene expression profiles

The application of phylogenetics methods to the analysis of gene expression profiles from individual tumors may not be its most natural usage, but does provide an alternative to other approaches for the classification of microarray or transcriptomic analyses, such as hierarchical clustering (Fig. 1C). In this utilization of phylogenetics, genes can be coded as the 'characters', expression levels can be coded as discrete 'character states', and individual samples can be the 'taxa' (Fig. 1C and D). In one such discretization, gene expression changes can be converted into discrete character states based on whether they are upregulated (1), downregulated (−1), or effectively unchanged (0). Application of a phylogenetically-based algorithm then produces one or more trees based on the similarities and differences in gene-expression profiles, putatively grouping (or classifying) cancer or disease tissues relative to normal tissue expression profiles (Fig. 1D). In traditional phylogenetics, a distantly-related taxon is used as an outgroup; however, in the analysis of cancer vs. normal tissue, the "outgroup" comprises either a mixture of normal tissues from representative individuals or, ideally, normal samples from the same individual as each tumor sample. A number of studies using maximum parsimony [52–54] and distance [55,56] algorithms on gene expression data have suggested that the methodology classifies tumors into monophyletic groupings compared to 'normal' tissue controls [57]. These analyses suggest that phylogenetic algorithms or algorithms developed from phylogenetic algorithms

Fig. 1. Phylogenetics reveals evolutionary relationships between states. A. Characteristics from various species under study can be transformed into a binary character state matrix. A species that is known to possess the ancestral state of the given characters (e.g. here, the lamprey) can be included as an "outgroup" as a means by which to polarize the resulting tree. B. An unrooted most parsimonious tree obtained by choosing the topology requiring the fewest number of character changes. C. The unrooted tree is converted to a rooted tree by assuming that jawed vertebrates share a more recent common ancestor than the most recent common ancestor (MRCA) of the entire group. C. As a cancer research tool, phylogenetic analyses can be used strictly as a clustering algorithm to segregate individual patients by their progression status. D. Samples are collected from individual patients, a matrix of characters is constructed using gene expression, mutation status, or some other information, and a phylogenetic tree is generated. E. In a more direct application of phylogenetic methods, they can be used to analyze phenotypic/genotypic heterogeneity within a patient or disease location. In this example, samples are collected at different sites to construct a matrix and tree of progression F. Depending on the question being asked, samples can be collected longitudinally or from neighboring areas of a tissue a single site (e.g. primary tumor and metastatic nodules) to reconstruct the evolutionary history of the disease progression.