In: Bioinformatics
Editor: Renu Tuteja

ISBN 978-1-62100-913-9

Chapter X

Cancer Phylogenetics: Computational Modeling of Tumor Evolution

Felix Bast*
Centre for Biosciences, Central University of Punjab,
Bathinda, Punjab, India

Abstract

The field of phylogenetics is one of the core areas of Bioinformatics which deals with computational methods to infer evolutionary heritage of organisms and genes. While phylogenetics has been extensively utilized in taxonomy and systematics of organisms, it is only very recently that the system started expanding to other fields- most importantly in cancer biology where it profoundly transformed our understanding of clonal evolution. Many of our findings in cancer phylogenetics credit to the fact that the tumor is not merely a collection of transformed cells with random mutation events; rather it is an evolving population.

Many of the facets underpinning modern evolutionary synthesis can be applied to classify cancers and track its progression from initiating somatic mutation to symptomatic neoplasm. It is now widely accepted that all sub-clones within cancer are phylogenetically related and probability of a particular sub-clone progressing into neoplasm depending upon its time of initiation and evolutionary fitness. Computational models of tumor evolution have also contributed in identifying common clades- "cancer sub-types"- associated with particular cancers in different patients that in turn helped in translating our understanding of oncogeny to the development of "targeted therapeutics"- rationally designed drugs that are molecularly targeted to particular sub-types. Advent of next generation ultra-deep genome sequencing technologies has been rapidly transforming the very landscape of cancer phyogenetics.

This chapter introduces the concept of cancer phylogenetics and reviews some of the recent advances in this field. This chapter also summarizes various phylogenetic approaches including distance matrix methods, parsimony, maximum likelihood,

* Email: felix.bast@gmail.com.

Bayesian methods and probabilistic inference that have potential applications in cancer research.

1. Introduction

Bioinformatics deals with use of statistical, mathematical and computational methods for processing biological information. Over the past half a century Bioinformatics have triggered a genomic revolution in which algorithms were designed for applications such as DNA sequencing, genomic sequence analysis and microarray expression profiling that have greatly contributed in our understanding of cancer as a multifactorial evolutionary phenomenon.

Traditionally cancer research has concentrated on the discovery of oncogenes and tumor suppressor genes and it was only very recently that the importance of factors contributing in the disruption of normal cellular differentiation became apparent.

Most cancer therapeutics work only for a small subset of patients and this remains the same problem even for advanced molecularly targeted drugs, for the reasons unknown as of now. This problem is causing such a financial burden in our healthcare system wherein large proportion of population receiving expensive-however ineffective-treatment. Biological characterization of each patient's tumor is needed in order to develop a scenario for delivering right targeted therapeutics to right cancer type[1].

Many cancers consist of heterogeneous population of subclones containing over-expressed, mutated and silenced genes. Identifying an effective molecular target remains a serious problem due to this heterogeneity [2]. Currently there is a substantial interest in the field of targeted therapeutics in which molecularly designed drugs attack a particular abnormality or disease mechanism present in distinct subtype of cancer [3].

Proper identification and characterization of such subtypes are essential for this approach. The field of Bioinformatics has immensely assisted in the logical groupings of target subtypes for a particular drug.It can however be stated that notion of cancer subtype as a static entity in the disease progression is a simplification; progression of neoplasm is rather a dynamic process. We need to understand not only general pathways of disease progression but also specific molecular steps along the pathway in order to maximize our chances of successfully curing it [4].

Resisting subclones evade in treatment targeting overexpressing and mutated genes [5]. Targeting a key "red dot" gene -mutation of which in early oncogenesis leads to dysregulation of molecular pathway involved intumerigenesis of all subclones- seems to be the winning strategy [6, 7].

Examples include HER-2, EGFR, KIT and ABL in breast cancer, non-small cell lung cancer, gastrointestinal stromal tumors and chronic myeogenous leukemia, respectively. Identifying red-dot targets, development and application of drugs targeting these and molecular characterization of each patient's tumor demands development of a bioinformatics framework encompassing multidisciplinary teams focusing translational research [6].

2. Cancer as an Evolutionary Process

Neoplastic progression recapitulates evolutionary processes in many aspects. Cancer can be viewed as a large population of genetically and epigenetically heterogeneous cells competing amongst and against host cells for space and resources[8]. As aberrations on somatic cells have heritable effect on the fitness of clones, expansion or contraction of mutant clones follows some of the well characterized processes of evolution like the natural selection and genetic drift [9-11]. Evolution of the neoplastic clones (clonal evolution) normally selects for increased proliferation and survival, leading ultimately to invasion, metastasis and therapeutic resistance [12, 13]. Although consideration of cancer as an evolutionary and ecological process is not new, nothing much has been done for the application of evolutionary biology and ecology in the field of cancer research and therapeutics.

2.1. Mutations

Variations in populations need to be heritable for the evolution to happen. Degree of variability in the heterogeneous neoplasms in fact demonstrated to predict progression to malignancy; i.e., more genetically diverse the cancer population is, more easily it would progress to malignancy [2, 14, 15]. Genetic heterogeneity in return is a result of genetic instability and is a hall mark of all cancers studied till date[15-18].

While number of mutations clearly increases risk of a somatic cell to get transformed, most of the genetic (mutations) and epigenetic alterations observed in cancer are evolutionarily neutral; i.e., they do not much contribute to the fitness of the population. Most of the non-neutral mutations are thought to be deleterious and therefore act against neoplastic progression [19]. Number and nature of mutations needed for the transformation remain unknown even for the most widely studied cancer retinoblastoma. A thorough characterization of key transforming mutations will contribute in the development of biomarkers and identification of targets for cancer therapy[20-22]. Mutation rates themselves can be taken as a biomarker for neoplastic progression as well as to analyze drug efficacy.

2.2. Genetic Drift in Cancer

Genetic drift is a stochastic evolutionary phenomenon in which allele frequency of a gene changes in a population due to random sampling. It is thought to be a key process contributing in neoplastic transformation and progression.[10, 15, 23, 24] However most of the parameters that are needed to understand role of this process in cancer progression remains to be measured; for example, cell generation time, turn-over numbers and effective population size of cancer clones. Like elsewhere, genetic drift becomes more relevant in cancer clones when population size is low. Evolutionary events like population bottlenecks in which population size gets significantly reduced are demonstrated to be fertile grounds for genetic drift to take place. For example, loss of breast epithelium due to apoptosis in menstrual cycle is a normal process giving rise to the process of population bottleneck [25]. Significant reduction in population size in cancer therapies might also contribute in the

process of genetic drift [11]. Cells with fitness disadvantages can get fixed sometimes, especially after the cancer therapy, due to population bottleneck giving rise to the genetic drift. Selectively advantageous mutations and mutations that are evolutionary neutral- however linked to the previous (hitchhiker mutations) - gets fixated ultimately through this stochastic process[26]. As neutral mutations are thought to be occurring at constant rates, it can be utilized as a molecular tumor clock- a representative of time since the initiation of neoplasm [27, 28].

2.3. Natural and Artificial Selections in Cancer

Natural selection involvescertain individuals possessing adaptive traits getting fixed in a population through reproduction and isolation. Heritable variation of reproductive success is an invariable requirement for the natural selection and this happens frequently in neoplasms as the genetic alterations are heritable and confer selective advantages or disadvantages on the cell. Tissues with repeated cycles of wounding and proliferations apparently have the repercussions of natural selection to the max, as these cycles effectively sieves clones having the best survival strategies; i.e., most proliferating[9, 10]. The process by which an adaptive mutation spreading through population is known as selective sweep and this might happen in the case of neoplasm clones, typically ending in the fixation of mutations that increase its fitness[29-31]. Several issues of natural selection in cancer remain to be answered. For example, what makes mutation patterns so different between different types of cancers and different organs? Selective pressures acting upon different cell types/organs may be quite different and factors affecting those needs to be thoroughly understood. It is also interesting to know the fitness effects of mutations in the genes that are not being used in normal cell types.

Interventional therapies in cancer give rise to deliberate artificial selection upon the neoplasm, potentially resulting in the emergence of resistance clones. The Luria-Delbruck experiment demonstrated that genetic mutations do not arise in response to the selection in bacteria; selective pressure merely selects pre-existing mutations [32]. Applying the same principle in the process of carcinogenesis, it can be deduced that the earlier we intervene in cancer the less probable it would be for the emergence of resistant mutant. Cancers that result from relatively smaller number of mutations like retinoblastoma are genetically less heterogeneous and therefore less likely to harbor resistant mutants. Traditionally chemo-therapy has been administered as large pulsed doses at regular intervals. Taking clues from evolutionary biology it might be the case that low-but continuous-doses work better as it generate lesser resistant mutants.

3. Phylogenetics

Phylogenetics- an evolutionary classification system that relies exclusively on the genetic relatedness of the objects -has become an indispensable tool in post-genomic era. While phylogenetics has been extensively used in biological systematics, the system has only begun to get expanded to other fields such as gene finding, comparative genomics, haplotype

inference, cancer biology and diagnostics[22, 33-35]. Phylogenetic inferences have a universal assumption that evolution is a process of divergence that can be modeled statistically. While this first-order assumption works in most cases, there are several evolutionary processes that possess certain difficulties. For example, Horizontal Gene Transfer (HGT) and inter specific recombination are documented in many bacterial lineages that results in transfer of genetic material across distantly related species[36]. Accurate evolutionary relationship can be modeled in the presence of these evolutionary processes by a phylogenetic network.

3.1. Species Tree vs. Gene Tree

The basic phylogenetic inference connects species through a bifurcating tree (cladogram). A set of orthologous genes (loci) are sampled from different species and genealogies are estimated to construct this species tree. Internal nodes of these trees represent speciation and other major taxonomic events.On the other hand gene trees represent evolutionary history of individual genes and can provide information such as gene duplication, nucleotide substitution and extinction events [37]. As a gene at a locus in a genome replicates and copies are passed on to daughter cells, branching points are generated in the gene trees. Within a species tree many tangled gene trees can be found due to meiotic recombination. Different genes should form different trees; in fact, even a single locus could have many trees as a result of intragenic recombination [38-40]. In light of this incongruence between species and gene trees, how gene trees can be used to reconstruct species trees? This can be done under the assumption that time intervals between species-branching events are much greater than time intervals between lineage-branching events in each species, and therefore gene and species divergences are likely to be nearly concurrent.

3.2. Character Conflicts in Cladistics Analysis

Genetic relatedness-or similarity (homology)-between specimens are of two fundamental types; ancestral and derived. Ancestral homologies (plesiomorphy) are very misleading and dangerous for any type of phylogenetic inferences [41-43]. For example consider lineages of birds, crocodiles and lizards. Bird lineage has undergone rapid evolution to get wings and other skeletal adaptations for flight. In comparison, lineages of both lizards and crocodiles have evolved more slowly to retain ancestral reptile characteristics such as walking on four legs and scales, and left looking similar. This similarity between reptiles and crocodiles is ancestral homology and it does not provide evidence that they have a Most Recent Common Ancestor (MRCA)-also known as concestor-than either does with birds. Consider number of digits on the legs of frog, dog and horse. Frog and dog have five digits-an ancestral state for all tetrapods (group consisting of reptiles, amphibians, birds and mammals). Horses have only one of the five digits left. Thus inferring this ancestral homology to conclude that frog and dog have more recent common ancestor than either do with horse, we are fundamentally erred. Indeed both horse and dog are mammals and share an MRCA than either does with frog. However if we are studying relatedness of a frog, dog and fish, five digits on foot is no longer an ancestral homology as it was not present in the common ancestor of all three

species. This derived homology (synapomorphy) has evolved within the group we are studying and correctly tell us about evolutionary heritage (phylogeny).Derived homologies are reliable evidence that two species share MRCA. Phylogenetics is a quest to define groups of specimens by discovering these derived homologies (phylogenetic characters). Shared derived state among the group of evolving specimens, synapomorphy, is a potential "biomarker" for that group. This evolutionary definition of biomarker as synapomorphy offers a natural classification of cancers based on theirontogenicrelatedness [22].

Convergent evolution describes acquisition of same phenotype in unrelated genetic lineages. A classic example is the wing in birds and bats. MRCA for both birds and bats did not have wings; both groups acquired this trait via convergent evolution. The term homoplasy refers to the similarity between two taxa of unrelated ancestry due to convergent evolution. Similar to ancestral homologies, homoplasies are troublesome for the phylogenetic inference [44, 45].

Truly derived homologies cannot conflict and in an ideal scenario character conflicts can be reduced to zero by cladistics analyses. However in reality conflicts can only be resolved to something higher than zero because techniques are all prone to errors. Homoplasies can be mistaken for homologies and convergence can be deceptively exact. One way to deal with conflicting phylogeny is to infer the one supported by majority (majority consensus).

3.3. Coalescent Theory

Tracing the evolutionary tree backwards by discovering MRCAs of the specimens comes under the realms of genealogy [46]. Species genealogies typically utilize one or multiple independent genetic loci to delineate between species that are being analyzed. The coalescent theory deals with modeling genetic drift of a population back in time to investigate genealogy of antecedents. Therefore it is a retrospective model, attempting to trace all alleles of a gene shared by individuals to MRCA [47, 48]. The basic coalescent model is a random variant; takes into consideration of only genetic drift and assumes no natural selection, population structure, recombination and gene flow. However recent advances in this model allow several of the above mentioned extensions to be analyzed concurrently and have potential applications in the field such as disease gene mapping [47, 49, 50].

3.4. Molecular Phylogenetics

As large amount of nucleotide and amino acid sequences are made available through public domain repositories, range of biological topics that it influences through evolutionary consideration expands rapidly. Sequences of DNA and proteins can be used to infer phylogeny the way morphological characters used in the past.While proteins blazed the trail (sequencing Insulin and its phylogeny reconstruction [51]), most of the recent developments in the molecular phylogenetics are concerned with DNA sequences.

Homology/homoplasy distinction, while powerful in deciding which morphological characters are phylogenetically informative, is so much less powerful in the case of molecular phylogenetics, as there are only four distinct states for nucleic acids and 20 for proteins[42]. Functional convergences, such as evolution of wings in species that fly, fuels discovery of

phylogenetically informative morphological characters. If a definitive relationship between structure and function is not available, a conclusion is impossible. This is the case for molecules and therefore we have to treat them the way morphologists treat organs with unknown functions. At the same time quantities of evidence presented by molecules are very large in comparison with morphology. Only through molecular phylogenetic studies that we understand information regarding structure and function of unknown proteins from rapidly accumulating genomic sequence data.

4. Modeling Evolution

Nucleotide substitution models, which encode hypotheses on the relative rates of mutations along the DNA sequences, are crucial components of molecular phylogenetic analysis. This is because of the fact that the longer the amount of time two sequences are diverged from the common ancestor (*i.e.*, immediately after coalescence), it is more likely that two or more consecutive mutations occur on any particular nucleotide position. Genetic distances between two diverging sequences increase linearly only for a short time since the divergence. Once separated, probability for two mutations to occur at any particular site becomes higher; therefore simple distance matrix approaches will undercount number of mutation events. One particular problem is the issue of "long-branch attraction" in which two distantly related but convergently evolving sequences can be misinterpreted as closely related.

Mathematical models of sequence evolution include variables that represent features of the evolution, but the numerical values are not known *a priori*. These variables are termed as parameters. There are two statistical approaches in modeling molecular evolution. Empirical models derive values of parameters from pre-computed analysis of large dataset. Parameter values are estimated only once and are applied to all datasets and therefore result in fixed parameter values. In this kind of models, particular data will only have negligible influence. The alternative approach is parametric models where pre-specified parameter values are absent. Parametric models allow the parameter values to be derived from each dataset.

4.1. Models of DNA Evolution

Different parametric approaches were developed to model DNA evolution in which three main nucleic acid parameters were frequently taken; 1. Base frequency, 2.Base exchange-ability and 3.Rate heterogeneity.

Base frequency parameters gives frequencies of four nucleotides; A,T,G and C, averaged over all sites in a DNA sequence. Factors such as overall GC contents are known to affect base frequency parameters. These parameters also skew frequency of certain bases during substitutions. Base exchangeability parameters are concerned with relative frequencies of bases to be substituted for one another (A <> T, A <>G, A<>C, T<>G, T<>C and G<>C) minus frequency of individual bases. If two bases are biochemically similar, base exchangeability will be higher, so that similar bases get substituted for one another. For example, A and G are purines and C and T are pyrimidines and therefore frequencies of A<>G and C<>T (purine<>purine or pyrimidine<>pyrimidine; "transition substitutions") are

higher than that of the rest of the substitutions (purine<>pyrimidine; “transversion” substitutions). Describing each site’s rate of substitution as random sampling from a continuous probability distribution like ”gamma distribution” is a common approach to model rate heterogeneity and an important step to fit models to data. It is often used together with base exchangeability and base frequency parameters.First DNA substitution model (JC) was based upon base frequency parameters [52]. Two parameters, viz., base transitions and transversions were taken into consideration in subsequently designed Kimura-2 Parameter model -K2P[53], which was also based upon equal base frequency parameters. Felsenstein in 1981 introduced a model (F81) based up on unequal base frequencies in which substitution rate corresponds to equilibrium frequency of the target nucleotide[54]. Hasegawa, Kishino and Yano in 1985 unified last two models into a six parameter base exchangeability model (HKY) that distinguishes between transitions and transversions[55]. In 1986 a Generalized Time Reversible (GTR) model was developed by Simon Tavare in which six substitution rate parameters as well as four equilibrium base frequency parameters were taken into consideration [56]. Taken together with gamma distribution parameter, GTR is one of the most widely used DNA substitution model today.

4.2. Choosing the Right Model

Selection of an appropriate nucleotide substitution model is often critical for any phylogenetic analysis, especially given that under-parameterized or overly restrictive models produce aberrant behavior when assumptions are violated and over-parameterized or complex models are computationally expensive and may be over-fit. A common method to achieve this is by means of the pairwiseLikelihood Ratio Test (LRT) between the models that estimates “goodness of fit” between the model and the data set. One disadvantage in using LRT is that it always estimates higher likelihood scores for complex parameters than the simple. Therefore most of the algorithms implementing LRT, for *e.g.*, MODELTEST [57], are designed such that it selects the simplest model with likelihood scores not worse than more complex models. Another disadvantage for LRT is that as it is a pairwise analysis, order in which models are compared has a major effect on the one that is eventually selected. Newer model tests like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) calculates likelihood estimates on individual model than a pair so that these are independent on the order[58]. These tests also have inbuilt correction factors to penalize over-parameterized (in case of AIC) or complex (in case of BIC) models.

4.3. Complex Models of Evolution

Evolutionary models giving insights into structure and selection are extensively used in studies of amino acid sequence evolution. Although such models are very complex and demand superior computational capabilities, they provide important information about molecular evolution. Codon replacement models have been recently developed in studies of coding DNA sequence evolution, in which both DNA substitutions and evolutionary constraints acting upon protein product are concurrently taken into consideration[59]. For example, by studying relationships between rates of synonymous (nucleic acid substitutions

will not change Amino Acid) and non-synonymous (Amino Acid altering) DNA substitutions, models will be able to detect positive selection without *a priori* knowledge of protein structure or function.

Several other models are based in addressing association of heterogeneity of patterns and rates of evolution among sites with structural organization of RNA or protein by simultaneously considering phylogeny and protein structure[60]. Structural and functional properties of proteins were used to model amino acid sequence evolution by analyzing amino acid substitutions at various structural motives [61].

5. Overview of Inferential Methodology

Properties of molecular sequence data such as availability of large amount of characters and recognizability of independent characters have encouraged utilization of statistical models to infer phylogenies. Inference methods will ideally extract maximum amount of information available in the dataset, combine this information with prior knowledge of patterns of sequence evolution and will deal with model parameters whose values are not known *a priori*. Representation of molecular hypotheses about the evolutionary ancestry of the sequences is achieved by phylogenetic trees (phylograms). There are several statistical approaches exist to infer phylogeny from molecular sequences. In this section we will cover the broad two classes of statistical approaches that have dominated in the literature: distance matrix methods and discrete data methods, although neither of them reproduces the evolutionary tree absolutely.

5.1. Distance Matrix Methods

Distance matrix methods measure genetic distances between the sequences and therefore explicitly requires MSA as an input. Pairwise evolutionary distances of the sequences in MSA are calculated and represented in a rooted or unrootedphylogram such that closely related sequences appear in the same interior node. Advantages of the distance matrix methods lie in the analyses being fast and from its ability to model substitution bias to correct multiple mutations. It produces only one tree- seemingly the best bet- however, costing the phylogenetic accuracy at a great deal. Neighbor-Joining (NJ) algorithm is one of the widely implemented distance matrix methods [62]. Unlike other methods, NJ does not assume that the lineages concurrently evolve (molecular clock hypothesis) and therefore produces an unrooted tree. By including known taxa as outgroup, it is possible to root the NJ phylogram, and when done that way, it always produces an ultrametric tree (equal distance from root to the branch tips).

5.2. Discrete Data Methods

Two of the discrete data methods commonly used are Maximum Parsimony (MP) and Maximum Likelihood (ML). The MP, a relatively simple non-parametric statistical method,

infers that the best representation of evolutionary relationships is the one that requires minimum number of steps (*i.e.*, nucleic acid substitutions). The input data for MP analysis, known as "characters", are phenotypical or genealogical attributes that are heritable and observed to vary between the taxa. MP produces a number of phylograms with considerable topological variations and therefore an evaluation of all such phylograms is very complicated. A strict consensus phylogram is usually constructed by heuristic approaches that usually involve steepest-descent style minimization mechanisms. Major disadvantage of this method is that the character states are generally noisy to an extent that overly simplistic approach of MP results in erroneous conclusions. Its inability to apply nucleotide substitution models and the common notion of evolution being non-parsimonious, further limits MP's usefulness in phylogenetic inference.

Most of the modern phylogenetic analyses are based on ML, a parametric statistical method, which is reported to be more accurate (*i.e.*, more likely to predict the evolution) and robust (less sensitive to faulty assumptions and models) than other phylogenetic inferences [63]. ML criterion assesses probability of particular mutations by a substitution model and allows varying rates of evolution across both lineages and sites [54]. Highly probable ML phylograms tend to have Interior branches that require minimum number of mutations to construct, and *vice versa*. ML is a preferable method for phylogenetic analysis of distantly related sequences, although it demands greater computational capabilities.

A much faster alternative that is often simultaneously performed with ML in the same data set is Bayesian Inference (BI) - name of which was derived from an 18th century statistician Thomas Bayes. The BI combines prior probabilities of a phylogeny with likelihood of trees to produce posterior probability distribution on phylograms [64]. Because tree topologies and branch lengths are not treated as parameters-as in ML- but as random variables, it is impossible to obtain BI probabilities analytically. Therefore BI probabilities are approximated by numerical simulations like Markov Chain Monte Carlo (MCMC) or Metropolis Coupled MCMC (MCMCMC). These chains explore the posterior probability grids in an integrative manner with model parameters. Trees are then sampled at fixed intervals and a consensus tree is constructed. The proportion of time that the chain visited sampled trees having a particular interior branch of the consensus tree is expressed as Bayesian posterior probabilities (PP, [65]). The computer program Mr.Bayes is often used to estimate the BI [66].

6. Genetic Heterogeneity and Cancer Classification Systems

Cancer staging classifications were traditionally based on the degree of differentiation of the tumor sample as determined histopathologically. This is then statistically analyzed to assign a particular stage. Finding common basis for defining type and subtypes of cancer is possible only if each subtype has unique developmental and maintenance pathways.

In recent times it has become increasingly evident that disruption of normal differentiation is an important component of tumorigenesis [67, 68]. Discovery of genes involved in this disruption have started helping the researchers to map neoplastic differentiation pathways and to classify cancer subtypes according to their maturation status.

One method of studying role of a particular gene in an organism is by monitoring its expression level. Gene expression data have been widely used recently in classifying various cancer sybtypes[69-74].

Expression of a particular gene changes during cell development and level of its expression can be correlated with amount of mRNA it encodes. Expression patterns of many genes in parallel can be monitored by complimentary DNA (cDNA) microarrays[75, 76]. Microarray data shows that each cancer specimen has a unique set of genes that are being expressed at any particular time. Gene expression profiles of these specimens are unique to distinguish it from other specimens. If these values are averaged-as is the case with phonetic clustering based staging methods- profile's identity gets distorted.

Principles of phylogenetics were first applied in the field of cancer to aid in the classification. Several researchers have attempted to utilize hierarchical clustering methods that are primarily used in phylogenetics to organize genes discovered by microarray methods into phylogenetic trees based on their expression patterns. One such method is based upon pairwise average linkage cluster analysis [77]. Recently more advanced mathematical and statistical data mining approaches including geometric mixture models and neural networks have been successfully used for the construction of phylograms from microarray data (9-14) to aid in cancer classification [78-81].

7. Phylogenetic Structures in Cancer

Identification and characterization of novel cancer subtypes have reached new heights under the technical framework of molecular phylogenetics. The principle behind cancer phylogenetics is straight forward; cancers are not merely a collection of mutated aberrant cells; rather it is an evolving population. Ongoing evolution of the neoplasm –i.e., patterns of tumor progression- can be tracked through analyzing mutations since the transformation event.

Ultimatelyasubclone with best evolutionary fitness comes to dominate the picture. Development and clinical presentation of symptomatic neoplasm might appear as a linear process macroscopically. However in subcellular level, competing genotypes in heterogeneous subclones and various extinction events become apparent. There are various direct and indirect evidences for such cancer progression models [82].

Some cancer subtypes are observed to be having high prevalence of somatically acquired mutations [83, 84]. Immediately after applying a selective pressure like drugs molecularly targetedto a particular subclone, preexisting drug-resistant clones were observed to be rapidly emerging- a phenomenon reflected in modern evolutionary synthesis [85]. Long latency between initiating mutations and clinical presentation of neoplasm is yet another indirect evidence for these models [86].

More recently multiple genetically related subclones in a case of B-Cell Chronic Lymphocytic Leukemia and clonal interrelationships within were demonstrated by phylogenetic analyses aided by massively parallel *de novo* DNA sequencing methods [87]. These studies ultimately help in the identification of the subpopulations of invasive, drug-resistant and relapsing cells as well as annotating initiating genetic alterations.

7.1. Two Approaches

There are two approaches widely used in Cancer Phylogenetics. Tumor by tumor approach utilizes difference between tissues between tumor populations-inter-tumor heterogeneity. Cell by cell approach utilizes difference between individual cells within tumor-intra-tumor heterogeneity. These two approaches with various methodological avenues for discovering phylogenetic structures in cancer are illustrated in Figure 1. This can be compared with approaches used in phylogeography- a study deals with finding historical processes responsible for contemporary distribution of organisms. Sampling from one microenvironment helps to assess species diversity and distribution within that habitat. At the same time, sampling from a wider geographic area helps to discover inter-habitat heterogeneity and distribution patterns.

Based on the presence or absence of specific mutation events, microarray gene expression measurements and gene copy numbers measured by Comparative Genomic Hybridization (CGH), it was Desper et al. who applied the principles of phylogenetics for the first time in the field of cancer to calculate evolutionary distances for inferring inter-tumor *oncogenetic trees* [35].

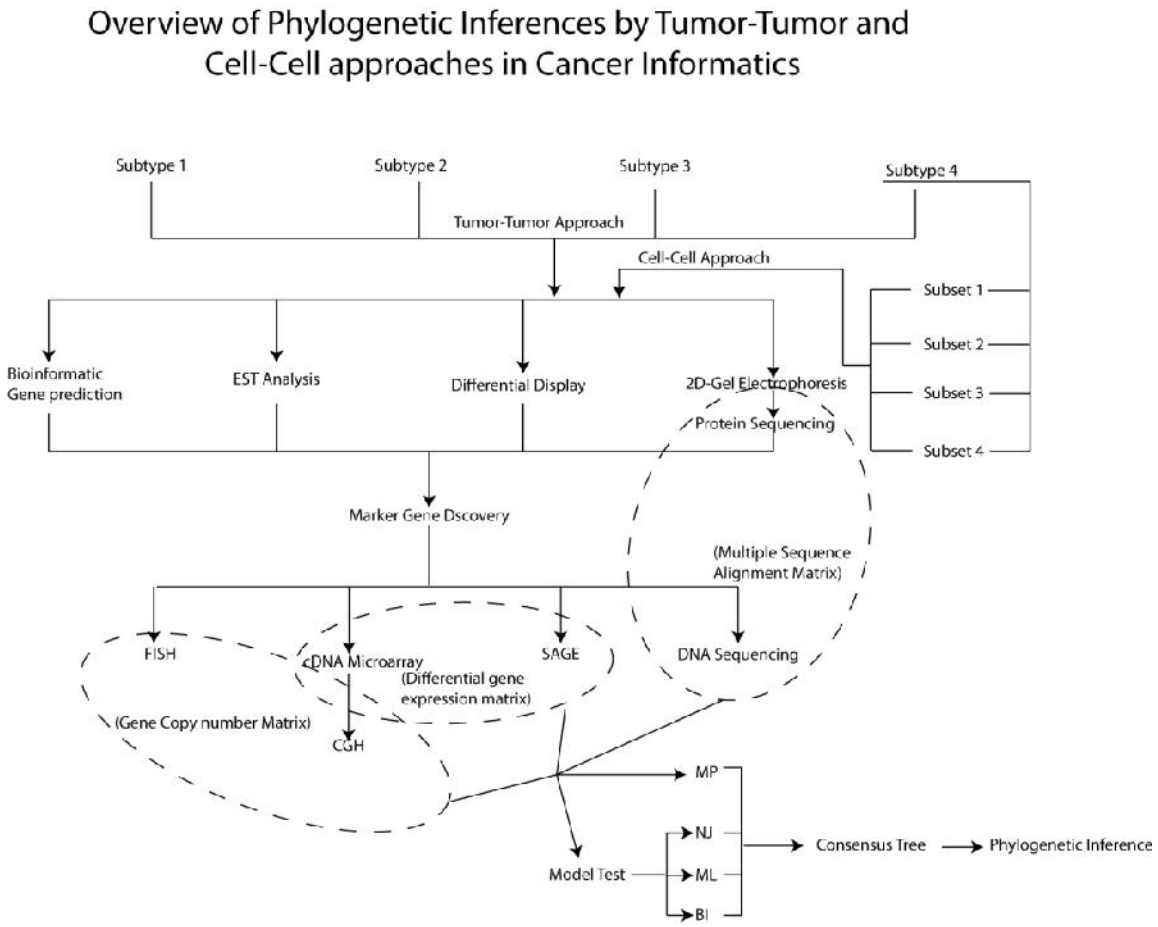

Figure 1. Two approaches and underlying methodological avenues for discovering phylogenetic structures in cancer. EST- Expressed Sequence Tags; SAGE- Serial Analysis of Gene Expression; FISH- Fluorescent In-situ Hybridization; CGH-Comparative Genomic Hybridization; MP-Maximum parsimony; NJ-Neighbor Joining; ML-Maximum Likelihood; BI-Bayesian Inference.

Average disease progression pathways evolving across a patient population were modeled by treating these cancer subtypes as internal nodes in the phylogenetic trees. Molecularly similar tumor subtypes form members of the same clades and distances of each subtype from the normal-state corresponds to the degrees of progression. Since then, many studies have conducted in this field and applied robust phylogenetic inference methods including maximum parsimony, maximum likelihood and Bayesian inferences.This tumor by tumor approach allows assays of many distinct probes per cancer capable of covering an entire transcriptome during expression profiling or copy number changes of a particular gene over an entire genome. However an accurate intracellular picture of cancer progression including existence of transitory cell populations and interrelationship between them cannot be drawn from this approach. It considers each tumor as a particular progression state in the overall disease progression and presumes that one can draw an evolutionary tree connecting all these states. However, tumors do not comprise homogenous populations of cells; it in turn preserves remnants of disease progression. While cell proliferation rates were observed to be gradually progressing as the cancer progresses, cells of the later stages augment, not replace, cells in earlier stages. That is, looking at individual tumors one can draw overall picture of cancer itself (Figure 2).

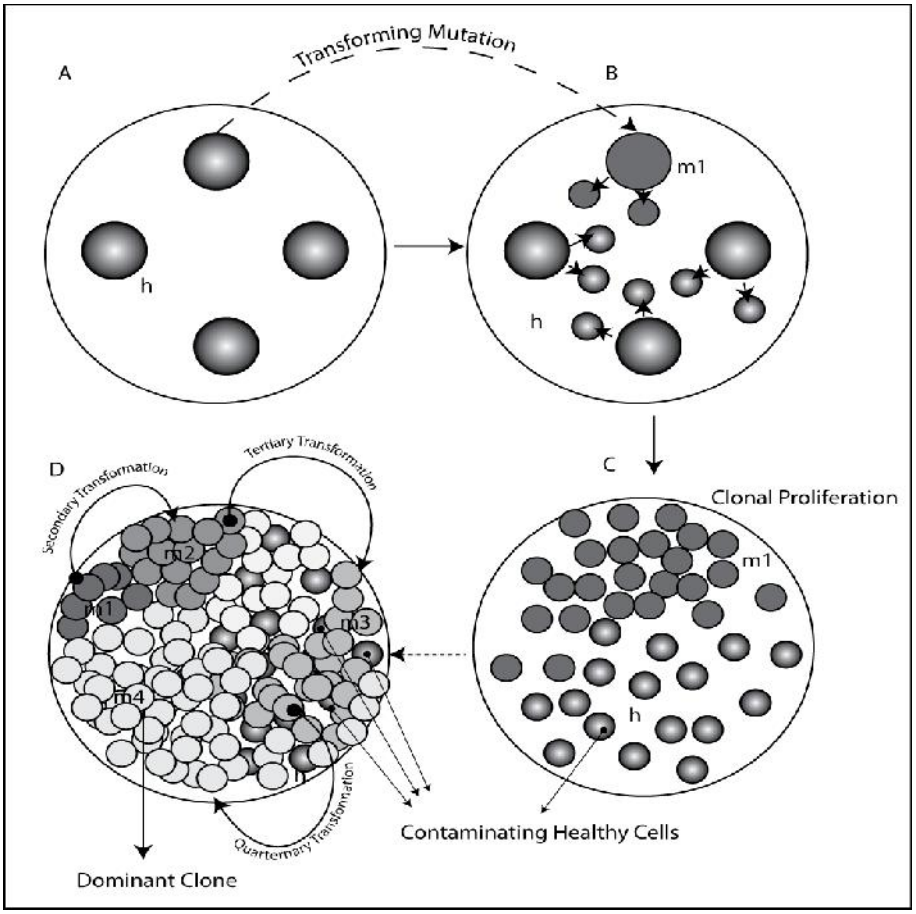

Figure 2. Neoplastic progression ("Chain Reaction") in cancer, leading to intra-tumor genetic heterogeneity. A. Healthy cells (h); B. Initial transformation event giving rise to mutated cell- a neoplasm (m1); C. Proliferation of M_1 neoplasm; D. Fully differentiated neoplasm retains the remnants of complete progression pathways.

This concept is the basis of alternative cell by cell approach which assumes that tumors preserve remnants of earlier cell populations as they develop [88, 89]. Tumors therefore contain transformed cells at various stages of progression as well as healthy "contaminating"cells. Various lines of evidence support this argument. Earlier studies involved gene copy number analyses by Fluorescent In-Situ Hybridization (FISH) that demonstrated single tumors as containing subsets of clones, each possessing one type of genetic alterations such as successive acquisition of a mutation sequence or differing degrees of gene amplification[90]. Evidences from the recent genomic sequencing experiments substantiate this finding that as tumors progress they retain remnants of ancient states along their progression pathways. Primary tumors, in comparison with secondary and other tumors resulted from metastases, shows higher degree of genetic heterogeneity indicating that metastases results from furtherhematopoietic differentiations of primary populations. By determining intra-tumor cellular heterogeneity it is possible to model common progression pathways and metastases.Since only a few previously characterized markers of progression can be used per cell at one time, this approach only gives relatively crude measures of progression states.

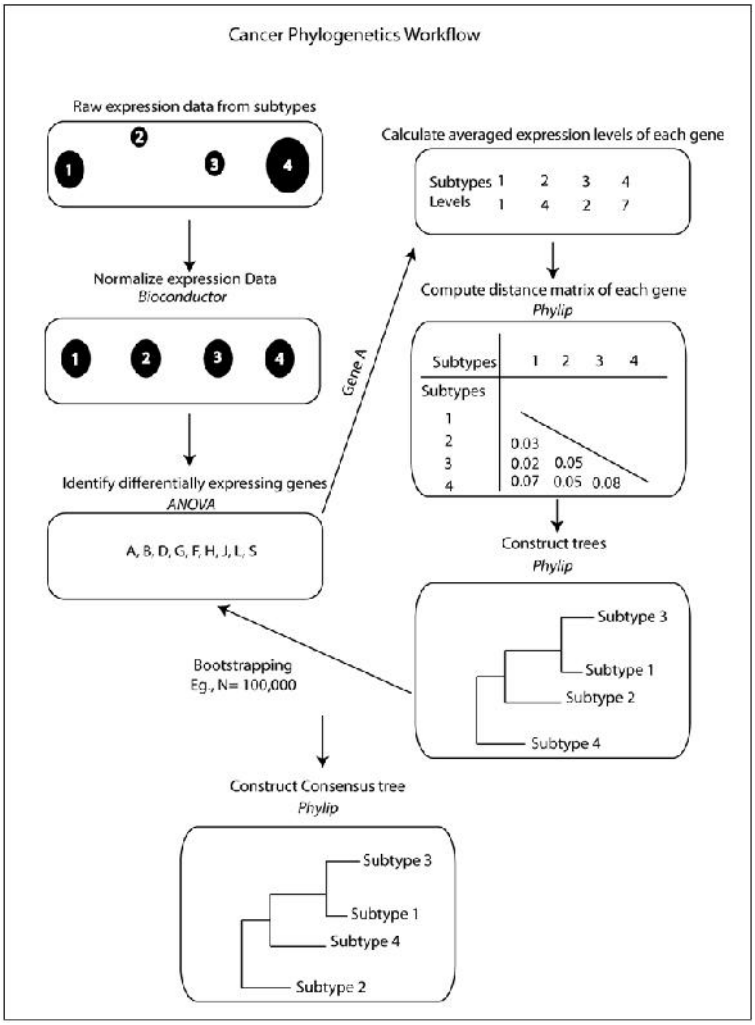

Figure 3. Cancer Phylogenetics workflow for tumor by tumor approach. Software/statistical parameters that can be used in main steps are indicated in italics. Modified from [95].

7.2. Outline of Methodology

Consider that we have microarray expression profile data from specimens that have been classified into different subtypes immuno/histo-pathologically. First step is to normalize the data across various specimens/subtypes. Software such as Bioconductor [91]are frequently used for the normalization. Next step is performing an ANOVA (one-way) to analyze for differentially expressed genes; i.e., a gene whose expression levels in one subtype differs from expression levels in at least one another subtype.Expression levels of each differentially expressed gene (discovered in previous step) is then averaged across all samples of each subtype to calculate average expression levels. Distance matrix is then computed from these averaged expression values, which in turn used to construct phylogenetic tree with software such as Geneious®(www.geneious.com) or Phylip®[92]. This process is then repeated (bootstrapping) with different sets of differently expressed genes to construct various trees with differing topologies [93]. A consensus tree is then constructed from these trees using appropriate algorithm such as Weighted Least Squares [94]. A workflow of this methodology is illustrated in Figure 3, which was successfully implemented for the phylogeny reconstruction of leukemia, breast cancer and liposarcoma subtypes[95].

Conclusion

Phylogeneticsis an important platform for cancer research by providing dynamic, predictive and seamless evolutionary classification and discovery of classes that accurately reveal biological processes and patterns between them. Analysis and interpretation of various phylogenetically informative data generated by modern cancer research will further facilitate our understanding of cancer as an evolutionary phenomenon. Since the advent of massively parallel next generation sequencing methods, huge quantity of omics data emerging at an ever increasing rate offer a unifying paradigm for the computational phylogenetics. Cancer phylogenetics will eventually aid in our quest for understanding differences in chemo-sensitivity among various clones and thereby to the realms of personalized medicine bycreating rationally designed and molecularly targeted drugs delivering right quantityto the right target group.

Future scenarios of cancer phylogenetics might include establishment of phylogenetic trees from genotyping distinct population of cancer cells isolated from spatiotemporally distinct parts of the tumor, including regions of invasion, metastasis and relapse. Single cell expression measurements may prove more valuable for cancer phylogenetics than measurements from gene or chromosome copy number changes. Identification of initial transformation event by phylogenetic inferences will also have potential clinical implications as the presence of residual cells having genetic alterations similar to initial transformation event could predict recurrence. Identification of initial transformation event can also facilitate the discovery of “red-gene” –chemopreventive and therapeutic targeting of which in the initial phases could ultimately prevent development of neoplasm altogether. In addition, oncogenetic trees reveal the directionality of change within a set of specimens, and could be of use in early diagnosis, prognosis, treatment assessment and biomarker identification.

References

[1] Bauvet, F., et al., [Therapeutic consequences of molecular biology advances in oncology]. *Bull Cancer*, 2009. 96(1): p. 59-71.

[2] Cooke, S.L., et al., Genomic analysis of genetic heterogeneity and evolution in high-grade serous ovarian carcinoma. *Oncogene*, 2010. 29(35): p. 4905-13.

[3] Wcislo, G., Molecular underpinnings of the targeted therapy for cancer. *Acta Pol. Pharm,* 2008. 65(6): p. 633-40.

[4] Dinh, P., C. Sotiriou, and M.J. Piccart, The evolution of treatment strategies: aiming at the target. *Breast,* 2007. 16 Suppl 2: p. S10-6.

[5] Turk, D. and G. Szakacs, Relevance of multidrug resistance in the age of targeted therapy. *Curr. Opin. Drug Discov Devel*., 2009. 12(2): p. 246-52.

[6] Gomez-Lopez, G. and A. Valencia, Bioinformatics and cancer research: building bridges for translational research. *Clin. Transl. Oncol.,* 2008. 10(2): p. 85-95.

[7] Colozza, M., et al., Breast cancer: achievements in adjuvant systemic therapies in the pre-genomic era. *Oncologist,* 2006. 11(2): p. 111-25.

[8] Axelrod, R., D.E. Axelrod, and K.J. Pienta, Evolution of cooperation among tumor cells. *Proc. Natl. Acad. Sci. USA*, 2006. 103(36): p. 13474-9.

[9] Maley, C.C. and B.J. Reid, Natural selection in neoplastic progression of Barrett's esophagus. *Semin. Cancer Biol*., 2005. 15(6): p. 474-83.

[10] Michor, F., et al., Somatic selection for and against cancer. *J. Theor. Biol*., 2003. 225(3): p. 377-82.

[11] Merlo, L.M., et al., Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer,* 2006. 6(12): p. 924-35.

[12] Takahashi, K., et al., Clonal and parallel evolution of primary lung cancers and their metastases revealed by molecular dissection of cancer cells. *Clin. Cancer Res*., 2007. 13(1): p. 111-20.

[13] Stilgenbauer, S., et al., Clonal evolution in chronic lymphocytic leukemia: acquisition of high-risk genomic aberrations associated with unmutated VH, resistance to therapy, and short survival. *Haematologica,* 2007. 92(9): p. 1242-5.

[14] Shackleton, M., et al., Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell*, 2009. 138(5): p. 822-9.

[15] Masramon, L., et al., Genetic instability and divergence of clonal populations in colon cancer cells in vitro. *J. Cell Sci.,* 2006. 119(Pt 8): p. 1477-82.

[16] Nowak, M.A., F. Michor, and Y. Iwasa, Genetic instability and clonal expansion. *J.Theor. Biol*., 2006. 241(1): p. 26-32.

[17] Maley, C.C., et al., Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet*, 2006. 38(4): p. 468-73.

[18] Breivik, J., The evolutionary origin of genetic instability in cancer development. *Semin.Cancer Biol*, 2005. 15(1): p. 51-60.

[19] Michor, F., et al., Stochastic elimination of cancer cells. *Proc. Biol. Sci*., 2003. 270(1528): p. 2017-24.

[20] Li, X., et al., Application of biomarkers in cancer risk management: evaluation from stochastic clonal evolutionary and dynamic system optimization points of view. *PLoS Comput. Biol*., 2011. 7(2): p. e1001087.

[21] Di Pietro, C., et al., The apoptotic machinery as a biological complex system: analysis of its omics and evolution, identification of candidate genes for fourteen major types of cancer, and experimental validation in CML and neuroblastoma. *BMC Med. Genomics*, 2009. 2: p. 20.
[22] Abu-Asab, M., M. Chaouchi, and H. Amri, Evolutionary medicine: A meaningful connection between omics, disease, and treatment. *Proteomics Clin. Appl*., 2008. 2(2): p. 122-134.
[23] Peng, B., C.I. Amos, and M. Kimmel, Forward-time simulations of human populations with complex diseases. *PLoS Genet*, 2007. 3(3): p. e47.
[24] O'Brien, S.J., A role for molecular genetics in biological conservation. *Proc. Natl.Acad.Sci. USA*, 1994. 91(13): p. 5748-55.
[25] Strassmann, B.I., Menstrual cycling and breast cancer: an evolutionary perspective. *J.Womens Health*, 1999. 8(2): p. 193-202.
[26] Bozic, I., et al., Accumulation of driver and passenger mutations during tumor progression.*Proc. Natl. Acad. Sci. USA*, 2010. 107(43): p. 18545-50.
[27] Shibata, D., Molecular tumour clocks. *Ann. Med.,* 1997. 29(1): p. 5-7.
[28] Shibata, D., et al., Somatic microsatellite mutations as molecular tumor clocks. *Nat.Med*., 1996. 2(6): p. 676-81.
[29] Maley, C.C., et al., Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. *Cancer Res*, 2004. 64(10): p. 3414-27.
[30] Akey, J.M., et al., Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol*., 2004. 2(10): p. e286.
[31] Zhang, J. and H.F. Rosenberg, Diversifying selection of the tumor-growth promoter angiogenin in primate evolution. *Mol. Biol. Evol*., 2002. 19(4): p. 438-45.
[32] Luria, S.E. and M. Delbruck, Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics*, 1943. 28(6): p. 491-511.
[33] Schwartz, R. and S.E. Shackney, Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics*, 2010. 11: p. 42.
[34] Shonkwiler, R.W., J.V. Herod, and Springer., Mathematical biologyan introduction with Maple and Matlab, 2009, Springer: New York ; London. p. xiii, 551 p.
[35] Desper, R., et al., Distance-based reconstruction of tree models for oncogenesis. *J.Comput. Biol*., 2000. 7(6): p. 789-803.
[36] Syvanen, M., Horizontal gene transfer: evidence and possible consequences. *Annu.Rev.Genet.,* 1994. 28: p. 237-61.
[37] Page, R.D. and M.A. Charleston, From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet Evol*, 1997. 7(2): p. 231-40.
[38] Liu, L. and D.K. Pearl, Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.,* 2007. 56(3): p. 504-14.
[39] Knowles, L.L. and P.B. Klimov, Estimating phylogenetic relationships despite discordant gene trees across loci: the species tree of a diverse species group of feather mites (Acari: Proctophyllodidae). *Parasitology*, 2011: p. 1-10.
[40] Zhang, L., From Gene Trees to Species Trees II: Species Tree Inference by Minimizing Deep Coalescence Events. *IEEE/ACM Trans Comput Biol Bioinform*, 2011.
[41] Meyer, A., Homology and homoplasy: the retention of genetic programmes. *Novartis Found Symp*, 1999. 222: p. 141-53; discussion 153-7.

[42] Wake, D.B., Homoplasy, homology and the problem of 'sameness' in biology. *Novartis Found Symp*, 1999. 222: p. 24-33; discussion 33-46.
[43] Rendall, D. and A. Di Fiore, Homoplasy, homology, and the perceived special status of behavior in evolution. *J. Hum. Evol.,* 2007. 52(5): p. 504-21.
[44] Yang, M. and G.J. Wyckoff, Detection of selection utilizing molecular phylogenetics: a possible approach. *Genetica,* 2011. 139(5): p. 639-48.
[45] Walsh, D.A. and A.K. Sharma, Molecular phylogenetics: testing evolutionary hypotheses. *Methods Mol. Biol*., 2009. 502: p. 131-68.
[46] Fu, Y.X. and W.H. Li, Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theor. Popul. Biol.*, 1999. 56(1): p. 1-10.
[47] Crandall, K.A. and A.R. Templeton, Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics,* 1993. 134(3): p. 959-69.
[48] Rosenberg, N.A. and M. Nordborg, Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet,* 2002. 3(5): p. 380-90.
[49] Tian, J.P. and X.S. Lin, The mutation process in colored coalescent theory. *Bull MathBiol.,* 2009. 71(8): p. 1873-89.
[50] Bulla, I., et al., HIV classification using the coalescent theory. *Bioinformatics*, 2010. 26(11): p. 1409-15.
[51] Sanger, F., Species differences in insulins. *Nature,* 1949. 164(4169): p. 529.
[52] Jukes, T.H. and C.R. Cantor, {*Evolution of protein molecules*}. 1969.
[53] Takahata, N. and M. Kimura, A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics*, 1981. 98(3): p. 641-57.
[54] Felsenstein, J., Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*., 1981. 17(6): p. 368-376.
[55] Hasegawa, M., H. Kishino, and T. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J.Mol. Evol*., 1985. 22(2): p. 160-174.
[56] Tavaré, S., Some probabilistic and statistical problems in the analysis of DNA sequences. *Some mathematical questions in biology-DNA sequence analysis*, 1986. 17: p. 57-86.
[57] Posada, D. and K.A. Crandall, Modeltest: testing the model of DNA substitution. *Bioinformatics*, 1998. 14(9): p. 817.
[58] Posada, D. and T.R. Buckley, Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.,* 2004. 53(5): p. 793.
[59] Yang, Z., et al., Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics,* 2000. 155(1): p. 431.
[60] Rzhetsky, A., Estimating substitution rates in ribosomal RNA genes. *Genetics*, 1995. 141(2): p. 771-83.
[61] Goldman, N., J.L. Thorne, and D.T. Jones, Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 1998. 149(1): p. 445-58.
[62] Saitou, N. and M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*., 1987. 4(4): p. 406-25.

[63] Huelsenbeck, J.P., The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol. Biol.Evol*., 1995. 12(5): p. 843.
[64] Huelsenbeck, J.P., et al., Bayesian inference of phylogeny and its impact on evolutionary biology. *Science,* 2001. 294(5550): p. 2310.
[65] Yang, Z. and B. Rannala, Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Mol. Biol. Evol*., 1997. 14(7): p. 717.
[66] Huelsenbeck, J.P. and F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 2001. 17(8): p. 754-5.
[67] Herynk, M.H., et al., Accelerated mammary maturation and differentiation, and delayed MMTVneu-induced tumorigenesis of K303R mutant ERalpha transgenic mice. *Oncogene*, 2009. 28(36): p. 3177-87.
[68] Stiewe, T., The p53 family in differentiation and tumorigenesis. *Nat. Rev. Cancer*, 2007. 7(3): p. 165-8.
[69] Fu, L.M. and C.S. Fu-Liu, Multi-class cancer subtype classification based on gene expression signatures with reliability analysis. *FEBS Lett*, 2004. 561(1-3): p. 186-90.
[70] Dyrskjot, L., Classification of bladder cancer by microarray expression profiling: towards a general clinical use of microarrays in cancer diagnostics. *Expert Rev. Mol.Diagn,* 2003. 3(5): p. 635-47.
[71] Sotiriou, C., et al., Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. USA*, 2003. 100(18): p. 10393-8.
[72] Ahr, A., et al., Molecular classification of breast cancer patients by gene expression profiling. *J. Pathol.,* 2001. 195(3): p. 312-20.
[73] Anbazhagan, R., et al., Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles. *Cancer Res*, 1999. 59(20): p. 5119-22.
[74] Golub, T.R., et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999. 286(5439): p. 531-7.
[75] Romualdi, C., et al., Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Hum. Mol. Genet*, 2003. 12(8): p. 823-36.
[76] Schena, M., et al., Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* 1995. 270(5235): p. 467.
[77] Eisen, M.B., et al., Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 1998. 95(25): p. 14863-8.
[78] Huang, M.L., et al., Usage of Case-Based Reasoning, Neural Network and Adaptive Neuro-Fuzzy Inference System Classification Techniques in Breast Cancer Dataset Classification Diagnosis. *J. Med. Syst.,* 2010.
[79] Wang, S.L., et al., Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. *Comput. Biol. Med.,* 2010. 40(2): p. 179-89.
[80] Ubeyli, E.D., Adaptive neuro-fuzzy inference systems for automatic detection of breast cancer. *J. Med. Syst,* 2009. 33(5): p. 353-8.
[81] Kuo, S.J., et al., Classification of benign and malignant breast tumors using neural networks and three-dimensional power Doppler ultrasound. *Ultrasound Obstet Gynecol*, 2008. 32(1): p. 97-102.

[82] Hanahan, D., The hallmarks of cancer. *Cell*, 2000. 100(1): p. 57-70.
[83] Stephens, P.J., et al., Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 2009. 462(7276): p. 1005-10.
[84] Greenman, C., et al., Patterns of somatic mutation in human cancer genomes. *Nature,* 2007. 446(7132): p. 153-8.
[85] Roche-Lestienne, C., et al., Several types of mutations of the Abl gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment. *Blood,* 2002. 100(3): p. 1014.
[86] Hong, D., et al., Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. *Science,* 2008. 319(5861): p. 336.
[87] Campbell, P.J., et al., Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. USA*, 2008. 105(35): p. 13081-6.
[88] Pennington, G., et al., Reconstructing tumor phylogenies from heterogeneous single-cell data. *J. Bioinform. Comput. Biol.*, 2007. 5(2a): p. 407-27.
[89] Pennington, G., et al., Expectation-maximization method for reconstructing tumor phylogenies from single-cell data. *Comput. Syst. Bioinformatics Conf*, 2006: p. 371-80.
[90] Smith, C.A., et al., Correlations among p53, Her-2/neu, and ras overexpression and aneuploidy by multiparameter flow cytometry in human breast cancer: evidence for a common phenotypic evolutionary pattern in infiltrating ductal carcinomas. *Clin. CancerRes.*, 2000. 6(1): p. 112-26.
[91] Gentleman, R.C., et al., Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 2004. 5(10): p. R80.
[92] Felsenstein, J., {PHYLIP}: phylogenetic inference package, version 3.5 c. 1993.
[93] Felsenstein, J., Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 1985. 39(4): p. 783-791.
[94] Desper, R. and O. Gascuel, Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol. Biol. Evol.*, 2004. 21(3): p. 587.
[95] Riester, M., et al., A differentiation-based phylogeny of cancer subtypes. *PLoS Comput.Biol.*, 2010. 6(5): p. e1000777.