Unravelling The History of The Milky Way with Phylogenetic Methods

Bede Denham

Abstract

This paper utilises stellar abundance ratios to investigate the use of phylogenetic methods in galactic archaeology. Current work in the field primarily applies these methods to small, real datasets. This paper applies the phylogenetic method to a large, simulated dataset to better evaluate their performance. This paper aims to combine galactic archaeology with phylogenetic methods towards unravelling the history of our galaxy, the Milky Way. To achieve this, a distance-based methodology is developed, resulting in a phylogenetic tree that strongly reflects the ground truth of the simulated dataset. The phylogeny created shows strong trends of increasing time of formation and increasing birth radii as the tree is descended. This paper demonstrates the strength of the interdisciplinary collaboration between phylogenetics and galactic archaeology and its use in unravelling the history of the Milky Way.

1. Introduction

The Big Bang is the current most popular theory in physics, which postulates that the universe began approximately 13.8 billion years ago. Soon after this, galaxies began to form (Ikpendu and Shinge, 2020). Our galaxy, the Milky Way, is estimated to contain over 100 billion stars. The Milky Way is a spiral galaxy containing a "thick disk", "thin disk", "bulge", "bar" and "halo" (Robin et al., 2003). The bulge and bar are located at the centre of the Milky Way with the thin disk surrounding them, followed by the thick disk surround the thin disk. Placed around the thick disk is the halo. This arrangement of structures can be seen in Figure 1. Currently, little is understood about how these star structures initially formed or how they evolved over time (but current theories will be discussed in the background information section of this paper).

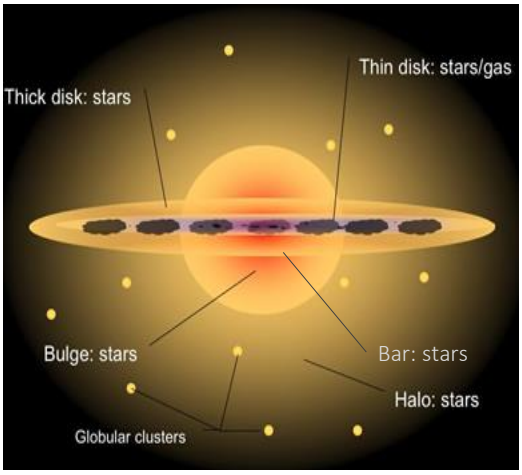

Figure 1 – Structural arrangement of the Milky Way. Source: Swinburne Astronomy, 2022.

Fortunately, modern astronomy allows individual stars in the Milky Way to be surveyed – giving access to a wealth of information. Using this information, analysis performed on the chemical compositions of the stars (along with their dynamics) provides insights into how these structures developed (Christlieb, 2002).

Phylogenetics is the study of evolutionary histories among species. Charles Darwin's seminal work "On the Origin of Species" in 1859 revolutionised the way life was viewed. In his Theory of Evolution, all species have evolved from one or few common ancestors (Darwin, 1859). A phylogeny displays the evolutionary relationship in a tree diagram. Phylogenies are "the basic structure necessary to think clearly about the differences between species, and to analyse those differences statistically" (Felsenstein, 2004). Typically, phylogenetic methods

use DNA sequence data and find a phylogenetic tree that "best" explains the observed data. The definition of "best" depends on the methods. For example, maximum parsimony methods define best as "requiring the least genetic change", whereas maximum likelihood methods define best as "statistically most likely" (Felsenstein, 2004).

This project aims to combine galactic archaeology with phylogenetic methods towards unravelling the history of our galaxy, the Milky Way. To do so, individual stars will be treated as "species" and a proxy for "evolutionary differences" between stars will be needed. In other words, data that approximates star "DNA" is essential for this project. Thankfully, due to the previously mentioned surveys of stars in the Milky Way, databases of stellar information exist. Most critical for this project is the chemical abundance data. Chemical abundance ratios provide insight into the formation of stars and can be treated as a proxy for the evolution of stars (Jofre et al., 2017; JØrgensen et al., 2020). This is because stars are formed from existing matters in the universe (matter in the Interstellar Medium, or ISM) which either came from the Big Bang or other stars. Further, over the period of their lifetime, stars produce energy via nuclear fusion – in which lighter nuclei are merged to form a single heavier nucleus (Haider, 2019). This process changes the chemical composition and hence, the chemical abundance ratios, of stars as they age. When these stars die, they release their material back into the ISM, enriching it and therefore, enriching the next generation of stars that are born from it. This is akin to DNA sequences as chemical evolution is also monotonic since the changing chemical composition accumulates over time. Whereas typical modern phylogenetic methods use nucleotide sequences to quantify DNA changes over time, this project utilises chemical abundance ratios to quantify stellar differences, which hopefully helps to elucidate the formation and evolution of the Milky Way. This paper will now discuss the formation of stars and the phylogenetic method in greater detail.

2. Background Information

2.1 Astronomy and Star Formation

The formation of the structures seen in the Milky Way is a hot research topic. Outlined below is a summary of what is currently known, along with leading theories. It is known that the thick disk formed in the first 3 billion years and contains stars that are mostly older than 10 billion years old (Gallart et al., 2019). These stars typically have lower abundances of more complex elements when compared to the thin disk. The thin disk formed after the thick disk and is composed of stars that are 6 billion years old on average (Gallart et al., 2019). The

formation of these disks (and the halo) is contended by two main theories, the two-infall model by Chiappini and Gratton; and a radial mixing model by Schönrich and Binney, supported by simulations by Kubryk et al. The two-infall model assumes that there were two infall episodes (where gas falls towards an astronomical body due to gravity) that assisted in the formation of the halo and thick disk. After the star formation burst following the first infall, a second infall leads to another star formation burst that forms the thin disk (Chiappini and Gratton, 1997). The radial mixing model states that disk structure of the Milky Way was instead formed via one infall episode and radial migration (Schönrich and Binney, 2009; Kubryk et al., 2015). Radial migration is defined as "the change in guiding centre radius of stars and gas caused by gains or losses of angular momentum that result from gravitational interaction with non-axisymmetric structure" (Grand et al., 2015). Essentially, this is the movement of stars caused by the gravity of other astronomical bodies. Hence, the radial mixing model contends that after the first infall, radial migration distributed matter and stars into the current formation (Schönrich and Binney, 2009; Kubryk et al., 2015).

The Big Bang released a massive amount of energy and matter as it expanded rapidly. When the universe was cool and dense enough for fusion (approximately 5 minutes after the Big Bang), all matter in the universe consisted of approximately 75% hydrogen and 25% helium (with traces of lithium) (Bertulani, 2019). The creation of these elements in this event is called Big Bang Nucleosynthesis (BBN) and from the interstellar medium (ISM) containing these elements, the first stars were formed. Over time, these stars created more complex elements such as oxygen, nitrogen, magnesium, and iron through stellar nucleosynthesis (Johnson, 2019). When these stars died, they released material back into the ISM. This "death" process was dependent on the type of star, see Figure 2.

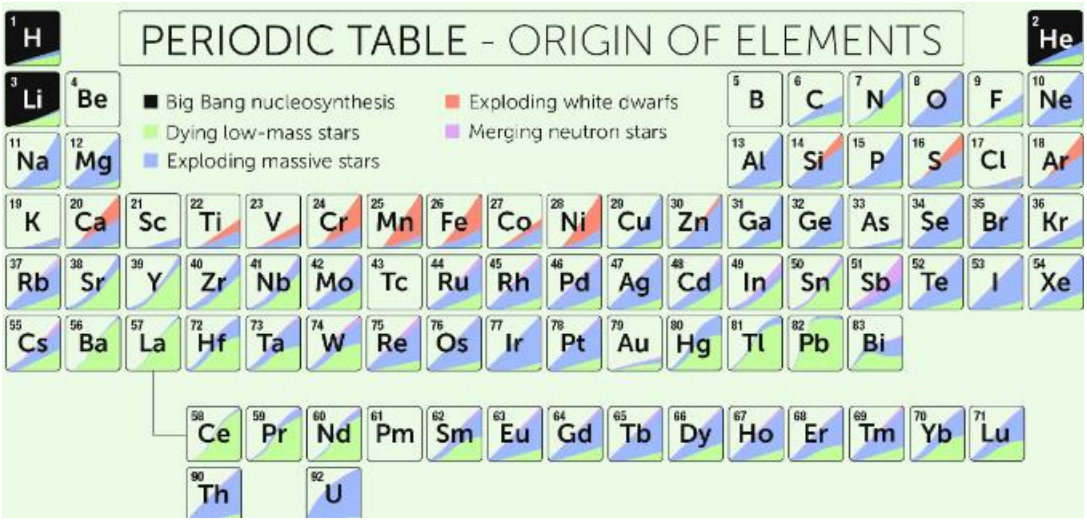

Figure 2 – Elements released by the different "deaths" of different types of stars. Source: Kobayashi et al., 2020.

Over time, new stars form from this material and begin again the creation of more complex elements via stellar nucleosynthesis. As this "life" cycle continues, to map the history of the Milky Way it is important that information regarding compositions of earlier stars is retained. Thankfully, less massive stars can have lifetimes in the hundreds of billions of years and can serve as fossil records for the chemical composition of stars over the timeline of the universe (Tolstoy et al., 2009). There are seven different types of main sequence stars: O, B, A, F, G, K and M (in descending size order). O types, the brightest and hottest type of main sequence star, are generally $50M_{\odot}$ (50 times the mass of the Sun) and live for 10 million years, whereas M types, the least bright and coolest type of main sequence star, are less massive (generally $0.2M_{\odot}$) and live for 200 billion years (Aguirre et al., 2013). These lifespans are forecast from the observed chemical processes occurring in the stars and as smaller stars burn fuel slower (leading to them being less bright and cooler), they tend to live longer (Aguirre et al, 2013). Further pertinent information regarding the formation history of the Milky Way is that due to the localised nature of material release upon death, stars that form near each other will be created from the same elemental deposits. Bland-Hawthorn et al. anticipate star clusters up to $10^5 M_{\odot}$, that is star clusters up to a one hundred thousand times the mass of the Sun, to be chemically homogeneous (Bland-Hawthorn et al., 2010). An additional effect of this localised material release is that different clusters of stars begin with different chemical compositions. This in turn leads to different chemical processes and stellar yield channels. Hence, the birth radius of stars is important in understanding the initial chemical composition of stars and consequently, the predicted chemical process. To clarify, the birth radius of a star in this paper pertains to its distance from the Galactic centre (its place of birth), not its size.

From this information it can be seen that there are two main drivers behind the chemical composition of stars: age and birth radius. The birth radius informs the starting chemical composition of a star and therefore knowledge of the starting point of its evolving chemical process. The age of a star informs what the composition of the ISM was like at the time of birth, which indicates the initial chemical composition of the star. However, radial migration complicates a direct inference using these two main drivers. One of the main benefits of applying the phylogenetic method to stellar heritability is that radial migration is effectively ignored. This is because after a star is formed its location has no effect on its chemical composition (Rojas, 2021). A direct genetic analogy can drawn to reinforce this – humans migrate, but their genetic composition is unchanged. A further advantage is that the selection function utilised for the stars is also of little importance to the final result. This is because, even though only a subsample of stars is being analysed (and hence branches of stars might be

missed), the validity of the tree topology is unaffected. Hence, the use of the phylogenetic method to approximate the evolution of stars and the formation of our galaxy by utilising chemical abundance ratios of stars as stellar DNA is supported.

The surveys that provide chemical abundance measurements (along with dynamics such as velocity and luminance) use high-resolution spectroscopy (Buder et al., 2019). The spectra probe used in these surveys measures the chemical composition of the photosphere, or exterior, of the stars (Reddy, 2019). Importantly, while the stars perform stellar nucleosynthesis and the interior of the star is enriched, the chemical composition of the exterior doesn't change (modulo some subtle effects) (Lambert, 2004). Therefore, what these surveys are effectively measuring is the chemistry of the ISM when (and where) each star was born. Hence, the measurement of a star that was born 5Gyr ago (5 billion years), will inform the composition of the ISM 5Gyrs ago – even though the chemical composition of the star itself will have changed significantly since its birth. Two leading surveys are the GALactic Archaeology with HERMES survey (GALAH) and Apache Point Observatory Galactic Evolution Experiment survey (APOGEE). The GALAH database contains stellar parameters and abundances of up to 23 elements for 342,682 stars while the APOGEE database contains 25 abundances for 437,485 stars (Buder, 2018; Queiroz et al., 2020). These two databases contain abundances for 37 unique elements. Stellar abundances are expressed as abundance ratios for usability, usually with respect to hydrogen or iron. Iron is preferred as when hydrogen is used correlations between ratios occur. These correlations arise because, for example, when a supernova occurs it produces all elements together (ie. Fe, Ni, Mn, etc.) (Wheeler and Sneden, 1989). Hence, when measured with respect to hydrogen they are highly correlated as opposed to when they are measured with respect to iron. However, the abundance ratio of iron is given with respect to hydrogen out of necessity. Further, they are expressed relative to the abundance ratios of the sun in logarithmic scale. The general form of abundance ratios is below in Equation 1, where X is any element (except Fe) and $\odot$ represents the sun:

$$\frac{X}{Fe} = \log\left(\frac{N_X}{N_{Fe}}\right) - \log\left(\frac{N_{X,\odot}}{N_{Fe,\odot}}\right)$$

Equation 1 – Abundance Ratio Formula

2.2 Phylogenetics and The Phylogenetic Method

The main problems in phylogenetics are the tracing of ancestry and evolution. By doing so, information regarding our origins and the origins of life on Earth can be inferred. Figure 3 displays a phylogeny that details the evolutionary history of humans and other species of apes and monkeys. In phylogenetic nomenclature, humans, chimpanzees and bonobos are "sister groups" meaning they share a common ancestor. Humans, chimpanzees, and gorillas form a "clade". A clade is a group of contemporary organisms sharing a common ancestor (Wilkinson et al., 2007). Further, trees can be created to display different relationships between "taxa". Taxa are the organisms at the tips of tree branches (eg. Humans). Different tree types include phenograms and cladograms, which vary how taxa are grouped. For example, phenograms structure the tree based upon steps of increasing similarity, while cladograms group taxa that share the most derived characteristics (Felsenstein, 2004). Phylograms and chronograms differ by meaning of branch lengths. For example, in Figure 3, the tree is a chronogram, whose branch lengths are proportional to time.

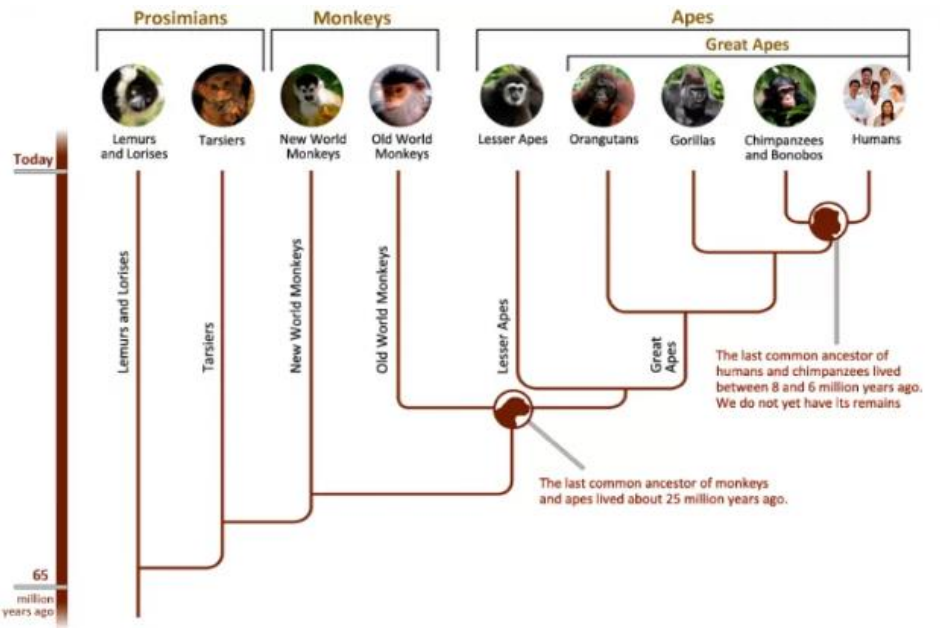

Figure 3 – Phylogeny detailing the evolutionary history of humans and other species of apes and monkeys. Source: Smithsonian, 2022

Several phylogenetic methods can be used to infer the trees. The four most common methods are maximum parsimony, maximum likelihood, Bayesian inference and distance matrix methods (Felsenstein, 2004). The general idea underpinning maximum parsimony is that the best evolutionary trees have "the minimum net amount of evolution" (Edwards and Cavalli-Sforza, 1964). Essentially, the best tree contains the fewest character state changes (or the least "events") necessary to validate the topology of the tree. To guarantee the best tree according to maximum parsimony, all possible trees must be assessed. This becomes

computationally expensive quickly for datasets with more species. Ernst Schröder computed the number of rooted bifurcating trees for *n* species is (2n-3)!. For n=9 species, there are just over 2 million possible trees (2,027,025), whereas at n=14 there are just under 8 trillion possible trees (7,905,853,580,625) (Schröder, 1870). Hence, parsimony methods usually rely on heuristic and branch and bound algorithms reducing the number of trees examined.

The other three phylogenetic methods are model-based, i.e., using a model of DNA sequence evolution, which is based on some assumptions. For example, the general time reversible (GTR) model assumes unequal substitution rates and base frequency (Felsenstein, 2004). To select appropriate models, likelihood ratio tests and Bayes factors are commonly used.

Maximum likelihood is a statistical method for estimating unknown parameters of a model. The maximum likelihood method finds trees having the highest conditional probability of observing a characteristic at the tips given the model (Felsenstein, 2004). The accuracy of the maximum likelihood methods is heavily reliant on the adequacy of the model. Maximum likelihood methods are even more computationally expensive than maximum parsimony. Therefore, the maximum likelihood methods also rely on heuristic algorithms for more efficient searching of the tree space.

Another statistically founded method is the Bayesian inference method. Specifically, the Bayesian method utilises the "posterior probability" of a tree as a judge of aptness (Felsenstein, 2004). The posterior probability is the conditional probability of a tree given the dataset. This is calculated using Bayes' theorem, explored in Equation 2 and Equation 3 below.

$$\Pr[A|B] = \frac{\Pr[B|A] * \Pr[A]}{\Pr[B]}$$

Equation 2 – Bayes' Theorem

For phylogenetic purposes we have:

$$\Pr[Tree|Data] = \frac{\Pr[Data|Tree] * \Pr[Tree]}{\Pr[Data]}$$

Equation 3 – Bayesian Inference

$$Where\ \Pr[Tree|Data] = Posterior\ probabilty,$$
$$\Pr[Data|Tree] = Likelihood,$$
$$\Pr[Tree] = Prior\ probability\ of\ phylogeny,\ \Pr[Data] = Evidence$$

Effectively, the posterior probability is the probability of the tree being correct. Again, this method is computationally expensive – even more so than maximum parsimony and maximum likelihood. Because of this, numerical methods are used to approximate the posterior probability distribution of the trees. Out of these the Markov Chain Monte Carlo (MCMC) numerical approximation method is the most widely used (Felsenstein, 2004). MCMC iterates through possible trees by assessing the effects of perturbations on the current tree. Each perturbation creates a new tree which is assigned a probability. Most often, trees that result in the highest probability are then selected and perturbed again. However, dependent on acceptance rate, trees with lower probability can be accepted (Larget and Simon, 1999). This process is repeated the posterior distribution becomes stable. However, it is worthy of note that chains can fail to converge.

Finally, distance matrix methods calculate the "genetic distance" between each pair of species and then construct a tree that mimics, as closely as possible, this observed set of pairwise distances (Felsenstein, 2004). The genetic distance is calculated by counting the number of character changes between two DNA sequences. It would seem that this method oversimplifies evolutionary complexity, however, simulations have shown that the loss of phylogenetic information is minimal and the results to be accurate (Felsenstein, 2004). However, distance-based methods are still less accurate than the maximum likelihood and Bayesian inference methods. Importantly, when using distance matrix methods, branch length represents the amount of genetic distance. Most distance matrix methods revolve around either the least squares criterion or the minimum evolution criterion (Pearson et al., 1999). The least squares criterion selects the best tree as the one where the discrepancy between observed and expected genetic distances is minimised. Whereas the minimum-evolution criterion, selects the best tree as the tree where the sum of the branch lengths is minimised (Pearson et al., 1999). Once a distance matrix has been created, trees are created through two main methods – the unweighted pair group method with arithmetic mean (UPGMA) method and the Neighbour-Joining (NJ) tree method.

Firstly, the UPGMA method was developed by Sokal and Michener in 1958. UPGMA is an agglomerative hierarchical clustering method that utilises the average linkage method (Felsenstein, 2004). Agglomerative clustering starts with each data point having its own cluster before matching and merging pairs of clusters that are the most similar (eg. have the least distance between them) (Chakrabarti et al., 2006). The average linkage method defines the distance between two clusters as the average of the distances between each pairing of data points contained in each cluster. The UPGMA method results in a rooted tree that assumes

equal rates of evolution with the same length from root to all tips. This requires the distances to be ultrametric, where the distance between any two points is equal to the length of the path connecting them (Felsenstein, 2004). Overall, the UPGMA method is simple and fast, however, it is unrealistic.

Secondly, the NJ algorithm was developed by Naruya Saitou and Masatoshi Nei in 1987. It is also an agglomerative hierarchical clustering method; however, it does not require ultrametric distances and utilises the star decomposition method. The star decomposition method is similar to the average linkage method whereby it begins with individual data points that are gradually clustered into pairs (Felsenstein, 2004). However, instead of using the average of the distances between pairs, the star decomposition method joins pairs such that the total length of the resulting tree is minimised. It is therefore an iterative clustering method that utilises a form of the minimum-evolution criterion. Further, the NJ algorithm allows unequal rates of evolution and results in an unrooted tree (Felsenstein, 2004). That is to say, the tree does not start at a datapoint, but rather a bifurcation from an arbitrarily set starting point. It is slower than the UPGMA method, however, produces more consistent results. Importantly, both of these distance matrix methods are significantly less computationally expensive than the maximum parsimony, maximum likelihood and Bayesian inference methods. They are better for application to large datasets and are ideal for application to continuous data. One notable disadvantage of distance matrix methods is the inevitable loss of information, however as aforementioned, simulations have shown the amount lost is remarkably small (Felsenstein, 2004).

2.3 Selecting The Right Phylogenetic Method For This Project

Comparing the above phylogenetic methods in the context of this project, a distance matrix method is best fit. There are two leading reasons for this. Firstly, there is data available on hundreds of thousands of stars and to not unreasonably limit the number of stars investigated in this project, a computationally efficient method is preferred. Secondly, the available star data is continuous and to avoid the loss of information involved in discretising the star data, a phylogenetic method that allows for continuous data is preferred. Hence, since distance matrix methods are substantially less computationally expensive than other phylogenetic methods and allow for continuous data, they will be utilised in this project.

Further, a choice must be made regarding which distance matrix method to use. Between the aforementioned UPGMA and NJ algorithm, NJ is preferred for the following reasons. Firstly, the NJ algorithm allows for unequal rates of evolution while UPGMA assumes equal

evolutionary rates. Due to the localised nature of material release upon star death and the interrelated variance in stellar yield channels, stellar nucleosynthesis can vary between stars – particularly stars that are born far apart. Therefore, the NJ algorithm is preferred as it allows for unequal evolution rates.

Secondly, UPGMA produces a rooted tree, whereas NJ results in an unrooted tree. As the "root" star is unknown, a rooted tree can cause misleading information. Therefore, again, the NJ algorithm is preferred.

Finally, NJ is the more reliable of the two methods, however, UPGMA is faster. While greater speed allows for more stars to be assessed, the cost in reliability outweighs this benefit. In fact, the speed difference between the two distance matrix methods is negligible in the scope of phylogenetic methods. Further, the information gained by assessing a higher number of stars (as allowed by the greater speed of UPGMA) is outweighed by the loss of information due to the reduced reliability. Therefore, this project will utilise a distance-based matrix method with the Neighbour-Joining algorithm to unravel the history of our galaxy, the Milky Way.

3.Literature Review

Phylogenetic methods have recently been applied in astronomy (Jofre et al., 2017; Jofre et al., 2021; Martínez-Marín et al, 2020; Fraix-Burnet et al., 2006). Jofre et al. have published two influential papers on the topic. Firstly, "Cosmic phylogeny: reconstructing the chemical history of the solar neighbourhood with an evolutionary tree" in 2017. This paper attempted to create a phylogenetic tree of solar twins using chemical abundance ratios. Jofre et al. utilised distance matrix methods and calculated the chemical distance matrix via the Euclidean distance method. From this chemical distance matrix, a minimum evolution tree was constructed using the NJ algorithm (Jofre et al., 2017). To assess the robustness of the final tree both Monte Carlo simulations and bootstrapping techniques were used. The Monte Carlo simulations were used to propagate uncertainties in the abundance ratio measurements, while bootstrapping is used to evaluate the probability that individual branches of the phylogenetic tree are correct. The best tree was chosen using a majority-rule consensus algorithm, which creates the tree from the most commonly occurring bifurcations in the posterior trees (Jofre et al., 2017). The final tree displayed relatively accurate clustering for the majority of the stars analysed. These stars were clustered into three distinct groups, namely, the thick disk, thin disk and a cluster proposed to have been created through a star

formation burst. The study concludes that phylogenetic methods offer a flexible, multivariate approach to galactic archaeology and were successful in in recovering stellar populations. Further, by analysing the chemical enrichment rates of the stars in the thin disk and comparing them to that of the thick disk, the paper concludes that the star formation rate in the thick disk is significantly faster than in the thin disk.

Secondly, and perhaps most influentially on this project, Jofre et al published "Using heritability of stellar chemistry to reveal the history of the Milky Way" in 2021. This paper extends upon the 2017 paper and demonstrates the suitability of the phylogenetic perspective in galactic archaeology, lending support to the interdisciplinary collaboration (Jofre et al., 2021). In particular, Jofre et al. highlight the main advantage of the phylogenetic approach to be the derivation of substantial information regarding shared histories of the stars from a relatively small star sample. This information is two-fold – firstly, the tree structure can inform on stellar processes. For example, in the 2017 paper, it is proposed that one of the groupings was formed via a star formation burst due to the birth radii of the stars and their current location relative to what we already know of star groupings (eg. Thick and thin disk). Further, timing of these processes/events can be inferred. Secondly, outliers and anomalies in the tree structures frequently warrant further investigation, leading towards a more holistic understanding of stellar processes. This 2021 paper again applies the phylogenetic method to solar twins using chemical abundance ratios. Jofre et al. continue their use of distance matrices and the NJ algorithm; however, they used the maximum-clade-credibility algorithm as opposed to the majority-rule consensus algorithm to assess the robustness of the tree (Jofre et al., 2021). This is justified as the majority-rule consensus algorithm is dependent on the cut-off threshold employed, which is the minimum percentage of trees the bifurcation must occur in to be used in the final tree. Further, branch lengths are hard to accurately estimate as the final tree is a mixture of posterior trees with varying topology. Finally, for the same reason, it is possible that the final tree may not represent a true phylogeny. The maximum-clade-credibility algorithm used in the 2021 paper has similar selection criteria, however, differs choice of final tree. The algorithm evaluates every clade across each sample tree and chooses the final tree as the sample tree with the highest overall support. Hence, this algorithm avoids the drawbacks described above that are inherent in the majority-rule consensus algorithm. However, the use of the majority-rules consensus algorithm is supported by O'Reilly and Donoghue in their paper analysing the efficacy of consensus tree methods. Comparing the maximum-clade-credibility and majority-rule consensus algorithms, the paper concludes that the majority-rule consensus algorithm contains a lower proportion of incorrect

nodes when creating a consensus tree from the same posterior trees (O'Reilly and Donoghue, 2018). Additionally, Jofre et al. perform truncation of branches that are shorter than triple the median uncertainty in the chemical abundance ratios. This is justified as the performance of the NJ algorithm is accurate when shortest branch length is a least twice the magnitude of the error values (Atteson, 1999). The truncation performed changes the tree type from bifurcating to a multifurcating tree, known as a polytomy. This choice is further validated as an unresolved branching pattern (as can result from a polytomy) is preferred from over one that is incorrect (Jofre et al., 2021). Additionally, these unresolved branches can lead to the discovery of stellar events (e.g., star formation bursts caused by supernova explosions or galactic collisions). The paper concludes that there is immense potential for the interdisciplinary collaboration of phylogenetics and galactic archaeology as the final tree successfully groups stars into the thick and thin disk that agrees well with other independent works and highlights potential stellar events for further investigation.

In 2020, Martínez-Marín et al. published a paper utilising phylogenetic methods to investigate galaxies in the Coma Cluster alongside galaxies in the field using chemical abundance ratios. Similarly to Jofre et al. 2017 and 2021, the paper utilises distance-based matrix methods alongside the NJ algorithm to infer phylogenies of the galaxies. Further, the chemical distance matrix found was created in the same way as in Jofre et al. 2017 (outlined above). Additionally, the robustness of the phylogenies was assessed utilising Monte Carlo simulations and bootstrapping, with the final tree being chosen according to the majority-rule consensus algorithm. The study concludes by lending support to the use of phylogenetic methods in galactic archaeology through its findings, but states further investigation is needed to verify the validity of the findings (Martínez-Marín et al., 2020). This is partially due to the large sample size of galaxies chosen.

Fraix-Burnet et al. published a paper in 2006 applying phylogenetic methods to dwarf galaxies in the Local Group. The paper assessed dwarf galaxies using 24 properties, 3 of which were chemical abundance ratios. Fraix-Burnet et al. employed the maximum parsimony methodology by discretising the properties into bins. They then used bootstrapping to evaluate the robustness of the tree. After analysis, the paper returned 1041 most parsimonious trees when including all dwarf galaxies investigated, however, found a fully resolved tree using a subsample of 14 galaxies. In support of the interdisciplinary collaboration which Fraix-Burnet et al. coined "astrocladistics" (for exclusively maximum parsimony methods), the strict consensus tree created from the 1041 most parsimonious trees and the fully resolved tree both exhibited the same property behaviours and were fully compatible – indicating the

same evolutionary history (Fraix-Burnet et al., 2006). The paper concludes that phylogenetics is not only applicable to, but a powerful tool for, understanding the formation and evolution of galaxies. This in turn lends support to the use of phylogenetic methods to infer stellar evolutionary history.

This paper builds upon the work done in the reviewed papers by using simulations instead of real data. This is because, with simulated date, the ground truth is known – allowing for better evaluation of the methodology used. Further, instead of trying to reconstruct known clusters, this paper intends to analyse a significantly greater number of stars (2 orders of magnitude greater) to better evaluate the construction of the overall trend of star formation. The ability to simulate star data on such a level has only become possible recently and hence, what this paper attempts has not been done before.

The papers reviewed in this section lend support to the underlying methodology of this paper (the use of phylogenetics in galactic archaeology). Further, most papers employed distance matrix methods with the NJ algorithm. This validates the choice made in the previous section of this paper to utilise this methodology and algorithm. Additionally, the use of bootstrapping and both the maximum-clade-credibility algorithm and majority-rule consensus algorithm were supported.

4.Materials and Methods

The simulated Milky Way star data used in this project was modelled by Chen et al., utilising a single infall model. The model is multi-zoned, utilised an ISM with a cold and warm phase, and along with the single infall of fresh gas, incorporates radial flow of gas, radial migration and major nucleosynthesis yield channels (Chen et al., 2022). The yield channels include asymptotic giant branch (AGB) winds, core-collapse supernova (CCSN) explosions and type 1a supernovae (SNe 1a). The star formation in the modelled is governed by the infall rate of fresh gas and the Kennicut Schmidt law (Chen et al., 2022). This law relates the surface gas density and star formation rate within a region (Jiang et al., 2022). Stars are formed from cold gas and release their material into the warm ISM upon death – this process progressively augments the cold ISM. The model created by Chen et al. abides by many observational constraints for the chemical evolution of the Milky Way and is hence sufficient for use in this project.

The simulated dataset contained data on 83 chemical abundance ratios for 21,592 stars over a timeline of 2.55 billion years, along with their birth radius and time of formation. However, since many of these elements are not measured by either APOGEE or GALAH this project disregards them – leaving 37 elements measured elements. Further, the first billion years was chosen to prove the concept. Therefore, the dataset investigated in this paper consists of 37 chemical abundance ratios, along with the birth radius and time of formation for 1,666 stars over a timeline of a billion years.

A principal component analysis (PCA) was performed on the simulated star data to gain preliminary insights. It is worthy of note that as the model underlying the simulation data does not include noise, the variance may be larger in real life. Hence, since the PCA measures the spread of elements, the percentage of variance able to be explained by running a PCA on the simulation data will likely be lower than in reality.

As outlined in the background information section of this paper and supported by the literature review conducted, this project will utilise a distance based phylogenetic method and the NJ algorithm to infer the evolutionary history of the Milky Way. Two popular distance matrix methods were compared for their performance: hamming distance and Euclidean distance. The Hamming distance and Euclidean distance were calculated as outlined in Equation 4 and Equation 5, respectively, below:

$$D_{i,j} = \sum_{k=1}^{N} \left| \left[\frac{X_k}{Fe}\right]_i - \left[\frac{X_k}{Fe}\right]_j \right|, where \frac{X_k}{Fe} \; is \; the \; abundance \; ratio \; of \; a \; star$$

Equation 4 – Hamming Distance

$$D_{i,j} = \sqrt{\sum_{k=1}^{N} \left(\left[\frac{X_k}{Fe}\right]_i - \left[\frac{X_k}{Fe}\right]_j \right)^2}, where \frac{X_k}{Fe} \; is \; the \; abundance \; ratio \; of \; a \; star$$

Equation 5 – Euclidean Distance

After comparison of output sample trees, the Euclidean distance method was selected. This is also supported by current literature (Jofre et al., 2017; Martínez-Marín et al., 2020; Jofre et al, 2021).

Further, the tree produced will be evaluated based on its robustness using bootstrapping techniques to create 100 sample distance matrices. This procedure is commonplace in phylogenetics and was employed in reviewed papers (Jofre et al., 2017; Martínez-Marín et al.,

2020; Jofre et al, 2021). To perform bootstrapping, the columns of elemental data were resampled with replacement. This was achieved by randomly selecting 37 columns (the number of abundance ratios used) and creating a new dataset with these columns. A distance matrix is then calculated using this new dataset. Effectively, this procedure randomly weights different elements to observe the effect on the output tree. To construct the trees, DecentTree was used as it provides highly optimised and parallel implementations of the NJ algorithm. Especially for large datasets (like the one used in this paper), it was proven by Wang et al. that DecentTree exhibits improved performance in comparison to other current software. This project utilises the RapidNJ algorithm, supplied by DecentTree, which employs branch-and-bound techniques for optimisation (Simonsen et al., 2010). Decenttree accepts distance matrices in phylip format as input and outputs phylogenetic trees in Newick format. Phylip and Newick are common phylogenetic alignment and tree text formats, respectively. By inputting the 100 bootstrapped distance matrices in Phylip format into DecentTree, 100 posterior trees in Newick format are output. The effects of the resampling are then observed by creating a final tree utilising Scikit-Bio's tree representation module. Within this module is a utility function that applies the majority-rules consensus algorithm, hence, creating a single consensus tree. This algorithm was used as, although it has drawbacks as identified in Jofre et al.,2021, it is firmly supported by other recent papers and is compatible with this project. (O'Reilly and Donoghue, 2018; Bryant, 2003). The tips of the consensus tree have an associated bootstrap support value. This value represents the proportion of resampled trees in which the same branching structure occurred – and hence, the robustness of the tree can be assessed.

Given the large amount of datapoints in the trees, visual representation is an issue. This project over comes the issue by extracting tip data and graphing this to show overall trends (as is the aim of this project). What is important for the tree representations to show are high-level trends as well as the overall structure and hence, the evolutionary paths. Additionally, the tree viewing program, FigTree, was used to display the structure of the tree. Importantly, as the NJ method constructs an arbitrarily rooted tree, FigTree allows the tree to be re-rooted to create a more accurate phylogeny.

To implement the methodology outlined, Python was used. All the code used in this project can be found, along with the simulated dataset used, at the following Gitlab Repository: https://gitlab.com/bedetdenham/astrogenetics/.

5. Results and Discussion

The results, and subsequent analysis of the results, will be presented in sections for clarity. Firstly, the results of the analysis performed on the simulated dataset will be discussed. Secondly, the overall tree structure will be presented. Thirdly, an analysis and representation of the tips of the trees. This will elucidate the overall trends of the data in relation to the tree structure.

5.1 Data Analysis

From the scree plot in Figure 4, it can be seen that the first dimension of the PCA explains 72% of the variance in the star data. Figure 5 breaks this contribution down into its different chemical elements. The elements with the most variance are Barium (Ba), Lanthanum (La) and Strontium (Sr). After investigation, this is expected as they are S-process elements from AGB winds that are more stochastic than supernovae. Therefore, it is expected for them to have a greater spread. As a consequence of the magnitude of their difference being larger, these elements will be more highly weighted in the distance matrix calculations.

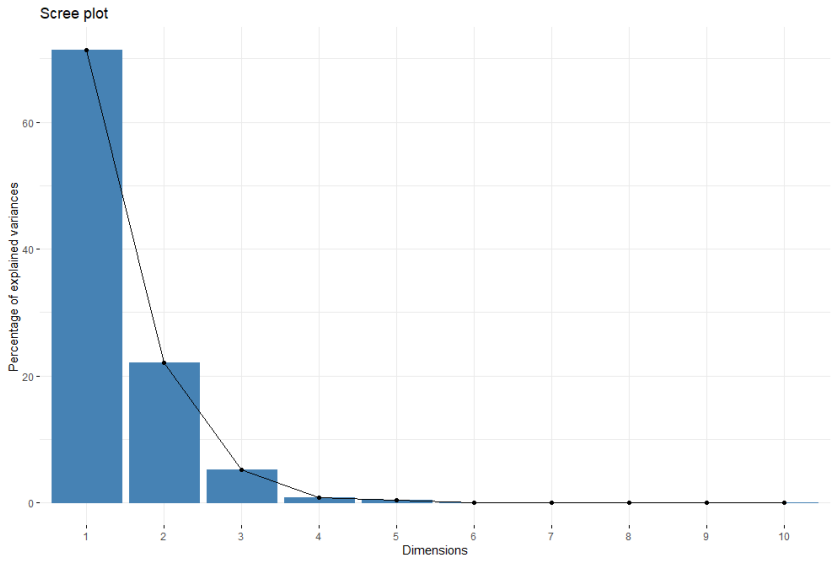

Figure 4 – Scree plot constructed from the simulated star data displaying the percentage of explained variance.

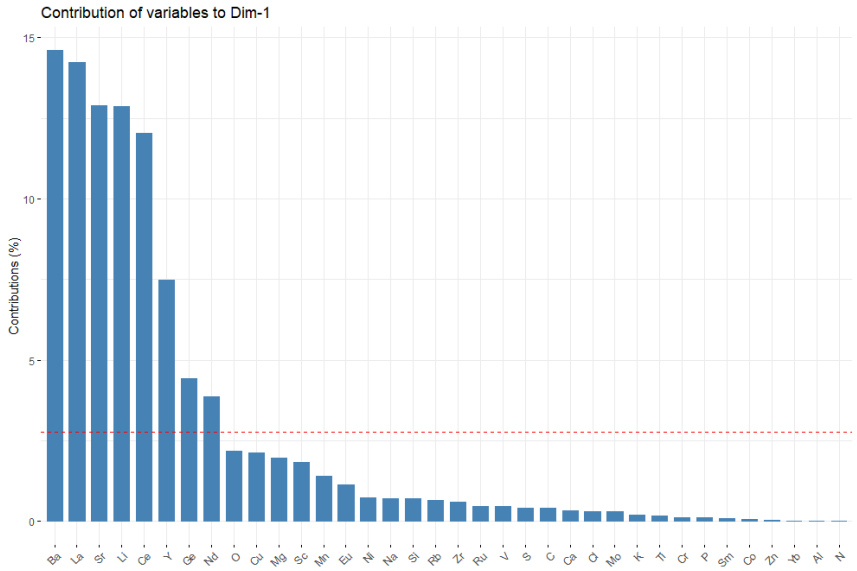

Figure 5 – Chemical element breakdown of the first dimension seen in Figure 4 above, which explains 72% of the variance in the simulated dataset.

Investigating the correlation of data to both confirm the results of the PCA above and to provide support to the model, the correlation was found between each abundance ratio and the time of formation as well as the birth radii. As expected from the PCA, Ba, La and Sr were the highest correlated with correlations of between 90% and 93% for all. Further, the correlations of the lighter elements such as Li, C and Mg were negatively significantly correlated (between -60% and -88%). This provides support to the underlying model that produced the simulation data, as this aligns with the theory. Further, the correlations between the birth radii and the elements are substantially weaker (none are greater than $\pm$30%) than those seen with the time of formation. This is indicative for two things, firstly, that age is the dominant variable in predicting the chemical composition of a star. Secondly, that the effects of radial migration are likely to be the cause of the substantially weaker predictive strength of birth radii. This highlights one of the benefits of the methodology being employed in this paper. That is that phylogenetic methods inspect only the chemistry and hence do not need to consider star dynamics such as radial migration – significantly reducing factors that commonly make inference difficult. Finally, there is a small (20%) correlation between the time of formation and birth radius of star. This aligns with the proofs that the galaxy is expanding and hence, provides support to the model used. A table of correlations can be found in Appendix A.

Figure 6 displays the expansion of the Galaxy. This can be seen in the bottom right corner, as for the first few hundred years no stars are born more than 11kpc from the Galactic centre. However, from approximately 400 million years onwards, stars begin to form there as the Galaxy expands. This provides support for the underlying model used for the simulation, as this reflects a fact that is known.

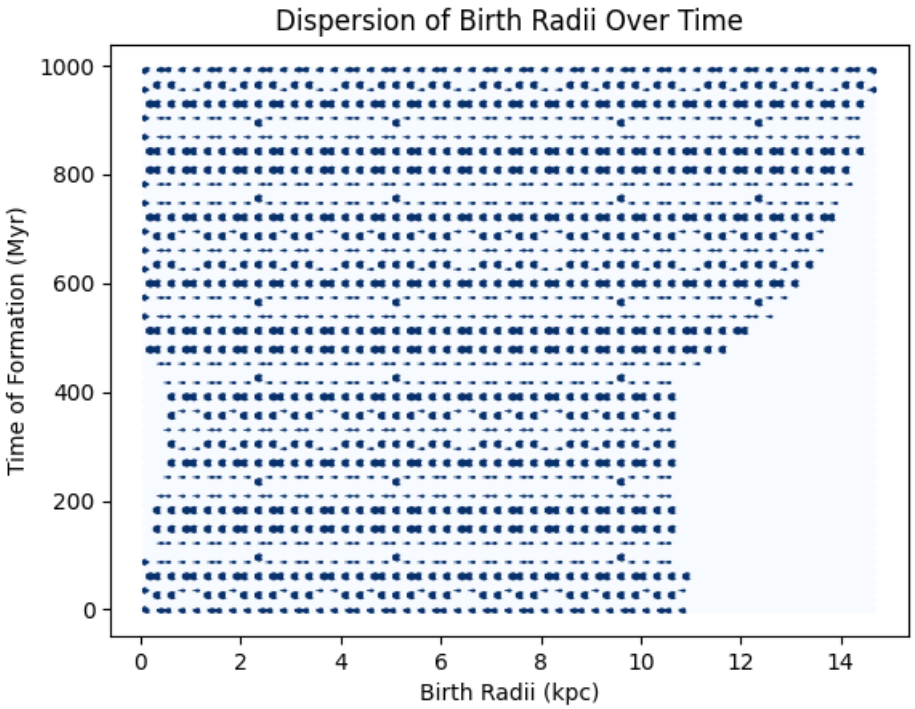

Figure 6 – Hexagonal bin plot of time of formation against birth radii. Depicts evidence of the simulated Milky Way expanding, which aligns with what is known about the real Milky Way.

5.2 Tree Structure

The final tree constructed with the full dataset, seen in Figure 7, can be seen to have three main branches. The starting point of these three branches are highlighted. This tree is colour-coded by birth radii and displays consistent grouping. This provides strong initial support for the performance of the methodology employed, as stars with similar birth radii are expected have similar stellar abundance ratios and hence, are expected to be clustered. Knowledge of the overall structure of the tree is essential to the analysis undertaken in the next section. Further, although a majority-rules consensus tree was created through bootstrapping, it is not used as the final tree. This is because the resultant consensus tree was unresolved, displaying primarily polytomies as a result of collapsed branches. This suggests a significant amount of uncertainty in the data as this would reduce the amount of agreement between trees, leading to branches being collapsed as they fall below the 50% threshold. As a consequence, the structures seen in Figure 7 cannot be seen. The tree can be seen in Appendix B. It is expected that the tree would be better resolved if more abundance ratios had been used.

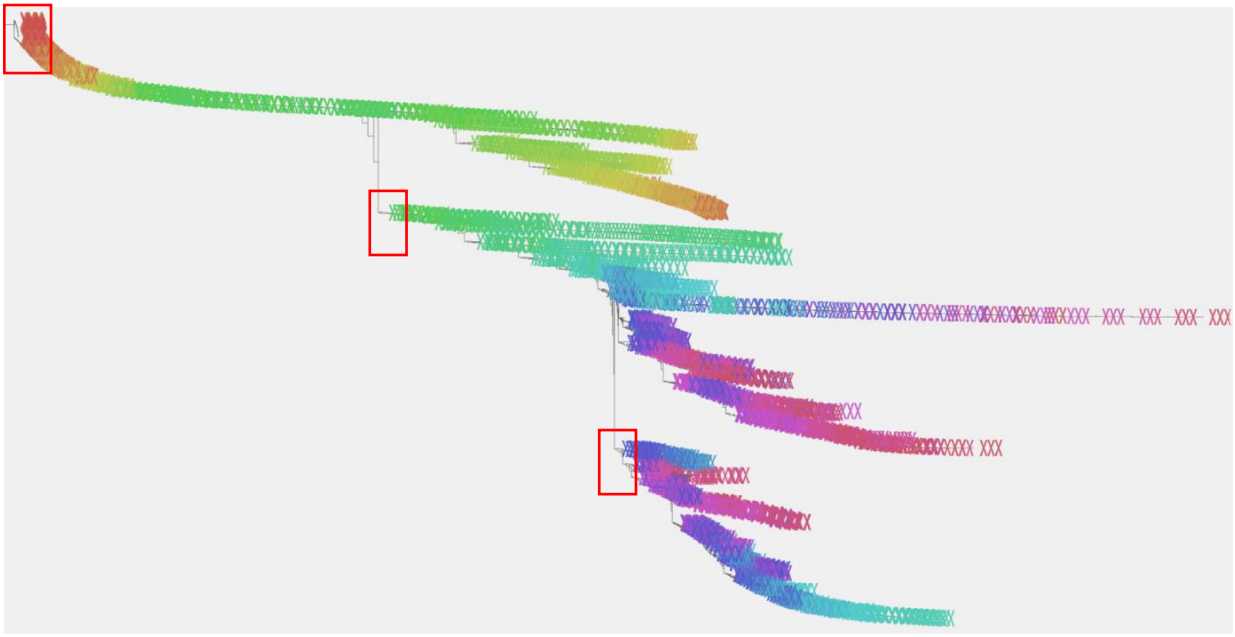

Figure 7 – Final re-rooted tree with three main branches. Tips are colour coded by birth radii and show consistent gradients.

Figure 8 displays the final tree of 15 randomly subsampled stars. Although the overall tree structure provides little insight on its own, with tip information, it reveals an important trend. The tips are named by time of formation/birth radii. It can be seen that the time of formation is dominant in the structure of the tree. Starting from the bottom right corner of the graph with the oldest star (0/10.625), the age of the stars displays a strong decreasing trend.

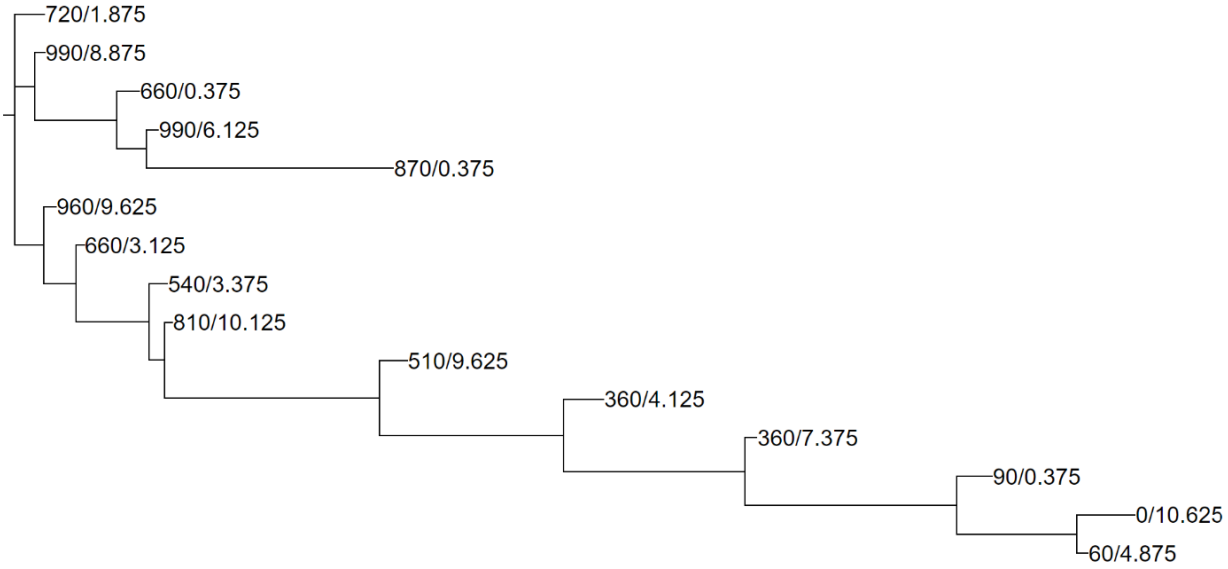

Figure 8 – Phylogeny of 15 randomly subsampled stars. Displays strong trend of increasing time of formation from right to left. This shows the bidirectionality of chemical distance and motivates the need for unfolding of branches.

Importantly, when analysing these trees, the bidirectional nature of the chemical distance must be taken into account. Hence, while the tree structure reflects the oldest star being a many-times-removed descendant this can in fact be inverted, making it the ancestor of the bottom branch. Therefore, this result accurately represents the ground truth of the simulated star data and gives strong support to the use of phylogenetic methods in galactic archaeology.

5.3 Tree Tip Analysis

Graphing the time of formation and birth radii of the stars against the number of ancestors they have in the tree will reveal overall trends in the formation of stars. The number of ancestors was calculated by the depth of the tip in the tree. As mentioned in the previous section, chemical distance is bidirectional and hence the number of ancestors had to be adjusted accordingly. This was done through manual inspection of the tree to find the first tip of each major branch. From this a subtree was created, which was then reversed and adjusted if necessary. Figure 9 displays the time of formation and birth radii against the number of ancestors for the entire dataset after this unfolding process was undertaken. Figure 9 shows two important trends that strongly align with the ground truth of the simulation data. Firstly, a trend of progressive increase in time of formation from ancestor to descendant. This trend of ancestors being older than descendants is an ideal result as this is, by definition, ancestry. Secondly, stars with higher birth radii (approximately greater than 11) tend to be older stars. As mentioned in Section 5.1, this agrees with the expansion of the universe. By aligning with the ground truth of the model used, along with what is currently known, strong support is provided by this result to the use of phylogenetic methods in galactic archaeology.

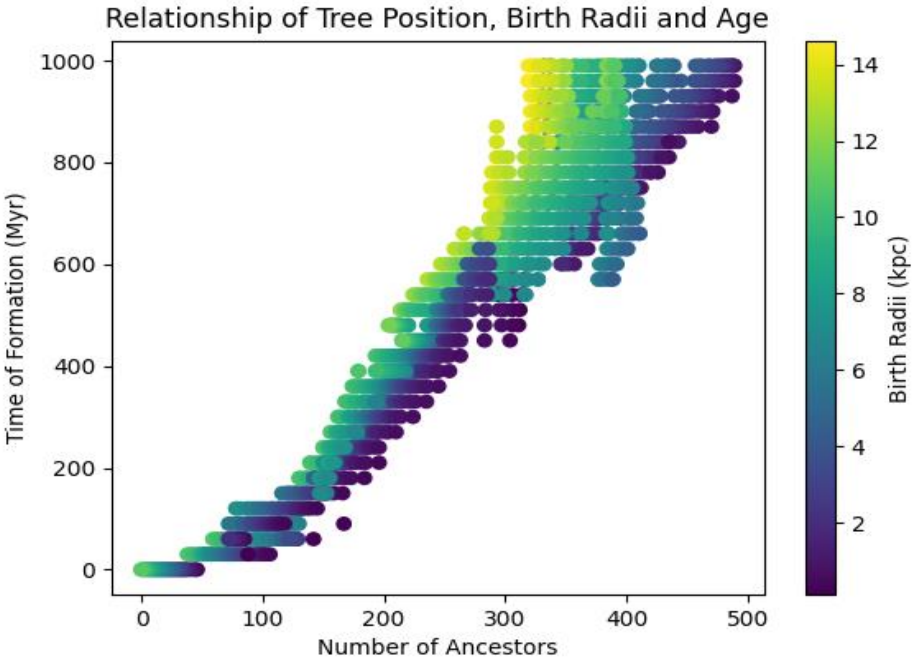

Figure 9 – Time of formation plotted against number of ancestors and coloured by birth radii. Displays strong trends of increasing number of ancestors and increasing birth radii over time.

6.Conclusion and Further Work

This paper presents a phylogenetic methodology that utilises stellar abundance ratios to create an evolutionary tree of a simulated Milky Way. This paper aimed to expand upon current research by applying phylogenetics to a large dataset created through a simulation of the evolution of the Milky Way. The simulated dataset contained 37 stellar abundance ratios, along with birth radii and times of formation, for 1,666 stars over a timeline of a billion years. Testing of the robustness of this simulated dataset was undertaken, with the results providing strong support for its validity. However, testing of the robustness of the tree indicated high levels of uncertainty in the data. Finally, the final tree created validates the proof of concept – that phylogenetic methods can be used to explore trends of star formation and consequently, to help reconstruct the history of the Milky Way. This was shown as strong trends were found between time of formation, birth radii and number of ancestors.

The results of this paper prompt further work, as the results demonstrate the huge potential of the interdisciplinary collaboration. For example, the methodology developed should be tested against a simulated dataset that involves multiple star formation events to test its performance different conditions. A further performance test would involve application of the methodology to a larger dataset that spans the entire history of the Milky Way (13.8 billion years), along with a wider selection of stellar abundance ratios. Finally, after performance testing under various conditions, a refined methodology should be applied to real star data to evaluate its performance against what is known about the formation of the Milky Way.

Acknowledgements

My sincere thanks to my supervisors, Minh Bui and Yuan-Sen Ting, for allowing me to get involved in such an exciting, frontier research area. Their guidance has been invaluable. Further, thank you to Boquan Erwin Chen for providing the simulated dataset used in this paper and to Trong Nhan Ly for his troubleshooting help. Additional thanks to James Barbetti, for developing DecentTree, but more importantly for spending his weekend assisting me in using it. Finally, a big thank you to Tony and Matilda Day for their endless support, both technical and otherwise.

7.References

Aguirre, V. S., Basu, S., Brandao, I., Christensen-Dalsgaard, J., Deheuvels, S., Doğan, G., Chaplin, W. (2013). Stellar ages and convective cores in field main-sequence stars: first asteroseismic application to two Kepler targets. The Astrophysical Journal, 769(2), 141.

Atteson, K. (1999). The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica, 25(2), 251-278.

Bertulani, C. (2019). Big bang nucleosynthesis and the lithium problem. Paper presented at the Journal of Physics: Conference Series.

Bland-Hawthorn, J., Krumholz, M. R., & Freeman, K. (2010). The long-term evolution of the galactic disk traced by dissolving star clusters. The Astrophysical Journal, 713(1), 166.

Bryant, D. (2003). A Classifica of Co sensus Methods for Phylogenetics. Paper presented at the Bioconsensus: DIMACS Working Group Meetings on Bioconsensus: October 25-26, 2000 and October 2-5, 2001, DIMACS Center.

Buder, S., Asplund, M., Duong, L., Kos, J., Lind, K., Ness, M. K., . . . De Silva, G. M. (2018). The GALAH Survey: second data release. Monthly Notices of the Royal Astronomical Society, 478(4), 4513-4552.

Buder, S., Lind, K., Ness, M. K., Asplund, M., Duong, L., Lin, J., . . . Bland-Hawthorn, J. (2019). The GALAH survey: An abundance, age, and kinematic inventory of the solar neighbourhood made with TGAS. Astronomy & Astrophysics, 624, A19.

Cavalli-Sforza, L. L., Barrai, I., & Edwards, A. W. (1964). Analysis of human evolution under random genetic drift. Paper presented at the Cold Spring Harbor symposia on quantitative biology.

Chakrabarti, D., Kumar, R., & Tomkins, A. (2006). Evolutionary clustering. Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.

Chen, B., Hayden, M. R., Sharma, S., Bland-Hawthorn, J., Kobayashi, C., & Karakas, A. I. (2022). Chemical Evolution with Radial Mixing Redux: Extending beyond the Solar Neighborhood. arXiv preprint arXiv:2204.11413.

Chiappini, C., & Gratton. (1997). The chemical evolution of the galaxy: the two-infall model. The Astrophysical Journal, 477(2), 765.

Christlieb, N., Bessell, M. S., Beers, T. C., Gustafsson, B., Korn, A., Barklem, P. S., Rossi, S. (2002). A stellar relic from the early Milky Way. Nature, 419(6910), 904-906.

Darwin, C. (2004). On the origin of species, 1859: Routledge.

Felsenstein, J. (2004). Inferring phylogenies (Vol. 2): Sinauer associates Sunderland, MA.

Fraix-Burnet, D., Choler, P., & Douzery, E. J. (2006). Towards a phylogenetic analysis of galaxy evolution: a case study with the dwarf galaxies of the local group. Astronomy & Astrophysics, 455(3), 845-851.

Gallart, C., Bernard, E. J., Brook, C. B., Ruiz-Lara, T., Cassisi, S., Hill, V., & Monelli, M. (2019). Uncovering the birth of the Milky Way through accurate stellar ages with Gaia. Nature Astronomy, 3(10), 932-939.

Grand, R. J., Kawata, D., & Cropper, M. (2015). Impact of radial migration on stellar and gas radial metallicity distribution. Monthly Notices of the Royal Astronomical Society, 447(4), 4018-4027.

Haider, Q. (2019). Nuclear fusion: holy grail of energy. In Nuclear Fusion-One Noble Goal and a Variety of Scientific and Technological Challenges: IntechOpen.

Ikpendu, E. L., & Ahmed, D. (2020). An Overview of the Cosmological Big Bang Theory of the Universe.

Jackson, H., Jofré, P., Yaxley, K., Das, P., de Brito Silva, D., & Foley, R. (2021). Using heritability of stellar chemistry to reveal the history of the Milky Way. Monthly Notices of the Royal Astronomical Society, 502(1), 32-47.

Jiang, B., Ciotti, L., Gan, Z., & Ostriker, J. (2022). Star formation inefficiency and Kennicutt-Schmidt laws in early-type galaxies. arXiv preprint arXiv:2208.03735.

Jofré, P., Das, P., Bertranpetit, J., & Foley, R. (2017). Cosmic phylogeny: reconstructing the chemical history of the solar neighbourhood with an evolutionary tree. Monthly Notices of the Royal Astronomical Society, 467(1), 1140-1153. doi:10.1093/mnras/stx075

Johnson, J. A. (2019). Populating the periodic table: Nucleosynthesis of the elements. Science, 363(6426), 474-478.

Jørgensen, J. K., Belloche, A., & Garrod, R. T. (2020). Astrochemistry during the formation of stars. Annual Review of Astronomy and Astrophysics, 58, 727-778.

Kobayashi, C., Karakas, A. I., & Lugaro, M. (2020). The origin of elements from carbon to uranium. The Astrophysical Journal, 900(2), 179.

Kubryk, M., Prantzos, N., & Athanassoula, E. (2015). Evolution of the Milky Way with radial motions of stars and gas-I. The solar neighbourhood and the thin and thick disks. Astronomy & Astrophysics, 580, A126.

Lambert, D. L. (2004). Observational Aspects Of Stellar Nucleosynthesis.

Larget, B., & Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular biology and evolution, 16(6), 750-759.

O'Reilly, J. E., & Donoghue, P. C. (2018). The efficacy of consensus tree methods for summarizing phylogenetic relationships from a posterior sample of trees estimated from morphological data. Systematic biology, 67(2), 354-362.

Pearson, W. R., Robins, G., & Zhang, T. (1999). Generalized neighbor-joining: more reliable phylogenetic tree reconstruction. Molecular biology and evolution, 16(6), 806-816.

Queiroz, A. B. d. A., Anders, F., Chiappini, C., Khalatyan, A., Santiago, B. X., Steinmetz, M., Barbuy, B. (2020). From the bulge to the outer disc: StarHorse stellar parameters, distances, and

extinctions for stars in APOGEE DR16 and other spectroscopic surveys. Astronomy & Astrophysics, 638, A76.

Reddy, B. E. (2019). Study of Lithium-rich giants with the GALAH spectroscopic survey. Monthly Notices of the Royal Astronomical Society, 484(2), 2000-2008.

Robin, A. C., Reylé, C., Derrière, S., & Picaud, S. (2003). A synthetic view on structure and evolution of the Milky Way. Astronomy & Astrophysics, 409(2), 523-540.

Rojas, F. E. (2021). The Consistency of Chemical Clocks Among Coeval Stars. Pontificia Universidad Catolica de Chile (Chile),

Schönrich, R., & Binney, J. (2009). Chemical evolution with radial mixing. Monthly Notices of the Royal Astronomical Society, 396(1), 203-222.

Schröder, E. (1870). Über unendlich viele Algorithmen zur Auflösung der Gleichungen. Mathematische Annalen, 2(2), 317-365.

Simonsen, M., Mailund, T., & Pedersen, C. N. (2010). Inference of large phylogenies using neighbour-joining. Paper presented at the International Joint Conference on Biomedical Engineering Systems and Technologies.

Smithsonian (2022, August 15). Genetics. Retrieved October 28, 2022, from https://humanorigins.si.edu/evidence/genetics

Swinburne Astronomy (2022). Thick disk: Cosmos. (n.d.). Retrieved October 28, 2022, from https://astronomy.swin.edu.au/cosmos/t/thick+disk

Tolstoy, E., Hill, V., & Tosi, M. (2009). Star-formation histories, abundances, and kinematics of dwarf galaxies in the Local Group. Annual Review of Astronomy and Astrophysics, 47, 371-425.

Wang, W., Barbetti, J., Wong, T., Thornlow, B., Corbett-Detig, R., Turakhia, Y., . . . Minh, B. Q. (2022). DecentTree: Scalable Neighbour-Joining for the Genomic Era. bioRxiv, 2022.2004.2010.487712. doi:10.1101/2022.04.10.487712

Wheeler, J. C., & Sneden, C. (1989). Abundance ratios as a function of metallicity. Annual Review of Astronomy and Astrophysics, 27(1), 279-349.

Wilkinson, M., McInerney, J. O., Hirt, R. P., Foster, P. G., & Embley, T. M. (2007). Of clades and clans: terms for phylogenetic relationships in unrooted trees. Trends in ecology & evolution, 22(3), 114-115.

Appendix A: Correlations

Element	Time of Formation	Birth Radii
Fe	85%	-28%
Li	-88%	23%
C	-60%	31%
N	78%	-3%
O	-46%	8%
Na	90%	-19%
Mg	-64%	10%
Al	86%	-19%
Si	58%	-6%
P	81%	-31%
S	79%	-11%
Cl	86%	-25%
K	89%	-18%
Ca	84%	-9%
Sc	90%	-18%
Ti	81%	-14%
V	-88%	8%
Cr	86%	-6%
Mn	33%	7%
Co	81%	-31%
Ni	-43%	-2%
Cu	90%	-17%
Zn	87%	-25%
Ge	90%	-16%
Rb	79%	-10%
Sr	91%	-9%
Y	90%	-12%
Zr	88%	-16%
Mo	75%	-15%
Ru	68%	-14%
Ba	93%	-4%
La	93%	-3%
Ce	93%	-3%
Nd	91%	-11%
Sm	86%	-16%
Eu	65%	-13%
Yb	89%	-15%
Birth Radii		19%

Appendix B: Consensus Tree