A framework for determining orthologues gene-drug targets based upon function, evolutionary selection rate, and physiological location.

By

Joshua M. Staley

B.S., University of Missouri – Kansas City, 2017

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Diagnostic Medicine/Pathobiology
College of Veterinary Medicine

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Gerald J. Wyckoff

Copyright

Abstract

The study of speciation, as a concept, existed prior to the elucidation of natural selection as the mechanism of evolution. However, speciation became a topic of scientific discourse starting primarily after the foundation of the modern synthesis. These disciplines operated independently; however, leading to the development of multiple theories of speciation. Some of the first efforts to consolidate speciation theory originated with the Biological Species Concept and the Modern Thesis, collectively introducing major factors of speciation (i.e. habitat, genealogy, evolutionary pressure, etc.) (1–3) To clearly define a species, a concise, reusable framework is urgently needed in the life sciences; specifically, translational and comparative medicine. Each of these fields depends on identifying suitable species to 1) develop a new treatment based on laboratory animal studies or 2) to take an existing treatment in humans and use it in animals. This is a major concern for pharmaceutical development research. The challenge of identifying a suitable species is complex due to a lack of a framework that allows for the comparison of species on the gene-gene level. Due to this complexity, gene-gene comparison is ignored which can lead to failure in a drug development trial. Subsequently, the cost to the development organization in both time and resources, and to society for lack of new and effective treatments, maybe great. With the proposed clearly defined, reusable framework, researchers would be able to find orthologues disease genes in animals ensuring that those genes are operating within the same cellular milieu and comparable physiology. For this reason, we have implemented a tool that calculates the evolutionary stability of the gene represented by the ratio $\frac{K_a}{K_s}$. This ratio can be used to find conserved genes across multiple species and assist in the determination of whether that gene is a good candidate for drug targeting and in which species the gene exists. In order to meet the needs of researchers in the field; we began integrating

GenBank, KEGG, and Refseq, with this tool, to allow those in the field to easily search across species.

Table of Contents

List of Figures

Acknowledgments

I would like to express my gratitude to the following teachers. To my major professor and mentor Dr. Gerald J. Wyckoff for his continued support and for introducing me to the field of evolutionary biology and Bioinformatics; without his guidance, I would not be where I am today. Thank you to the members of my committee: Dr. Jeffery Comer and Dr. Zhoumeng Lin for providing integral comments and feedback.

I would like to thank the following organizations that made the development of this framework possible: Canonical, for creating Ubuntu, an open-source development environment without which, experiments such as this would be prohibitively expensive; the Python Software Foundation for creating an easy-to-read and powerful programming language; BioNexus KC, for there finical support of both my research and to the life sciences across the animal health corridor; Johnson County Education and Research Triangle (JCERT), for there finical support; Bank of America, for there finical support.

I would not have been able to make it through each day without the day-to-day emotional support of Joe Rundquist, Tricia Jenkins, and Heather Woods. I would like to express my deepest gratitude to my parents, my two sisters and extended family, who provided me with finical and emotional support throughout my collegiate career. Finally, my fiancé Sarah, for her love, understanding, and for making my life an adventure.

Dedication

To my mom, for her example of strength and leadership and never-ending support and love for all that I do.

“THAT was a great game”

Chapter 1 - Literature Review

Pre-Darwinian Species Concept

Herbalists were among the first people to adopt a rudimentary classification system for living organisms dating back 5000 years to Mesopotamia. (4,5). Their classification system was exclusively for plants; specifically, those that could be used medicinally. Aristotle later postulated that there were "fixed forms" that existed in the surrounding universe; on which the universe has some effect; however, he did not know how to predict that effect or when it might be observed.(6,7) Epicurus, who studied under Plato, suggested theories about atoms, their properties, and their ability to affect other atoms; as well as go through physical changes over time.(8–10) Later his concepts of atoms were used to describe atoms as beneficial or harmful to other atoms, and humans.(9,10) Thereafter, humans have contemplated their order among other organisms and in doing so, stumbled upon the precarious task of grouping organisms into classes to be studied, now described as species. "Speciation, the evolution of reproductive isolation among populations, is continuous, complex, and involves multiple, interacting barriers"(11) and "any discussion of the genetics of speciation must begin with the observation that species are real entities in nature, not subjective human divisions of what is really a continuum among organisms."(12) Intuitively, over time, speciation researchers began by looking at the physical features of an organism and the field of phenetics was born; formally known as a Pre-Darwinian classification system, phenetics is based primarily on the phenotypic attributes of an organism.(4) Those working in the field operate under the Aristotelian assumption that all organisms are "fixed", in other words, they do not change over time. In doing this, two types of classification systems were employed: *priori* and *posteriori,* later described as artificial and natural, respectively. (4,13) Theophrastus, a student of Aristotle, was the first to classify plants

by habitat and whether they were cultivated or wild. For this reason, he is commonly referred to as the Father of Botany. (4,13) This is one of the earliest observed systems of taxonomy, which created operational taxonomic units (OTUs) using previously given criteria characterizing this system as *priori* classification. Linnaeus made this method famous in 1735 as is displayed in his *Regnum Animale* which established the binomial nomenclature system which is still used today and is fundamental to the field of taxonomy.(4,13,14) Antoine Laurent de Jussieu's *Genera Plantarum* introduced his "sexual system" in 1789, grouping plants according to the number stamens and pistils in the flower while employing Linnaeus' binomial nomenclature.(4,15,16) Conversely, he used many different characteristics and created OTUs as each characteristic was studied, naturally, characterized as *posteriori* classification.(4,12,13) Augustin Pyramus de Candolle employed this in his *Prodromus Systematis Naturalis Regni Vegetabili*, a 17-volume also known as Prodr. (DC.) pertaining primarily to dicotyledons. It has been used to develop the field of plant taxonomy through the 1900s.(4,17,18) Charles Darwin, prompted by Alfred Wallace, brought the discussion of evolution to the main stage. Its established implementation into speciation, however, did not come until later in the 20th century(7). Simpson standardized the nomenclature used when studying taxonomy by establishing fundamental terms (e.g. *Classification, Identification, Systematics, Taxonomy*) which created a scalable framework for the study of speciation, outlined in his text *Principles of Taxonomy*.(19,20)

Post-Darwinian Species Concepts

Through the 20th century, there was an explosion in the volume of OTUs that were characterized. This created a large amount of information to be studied. It became cumbersome and expensive to continue studying OTUs in a nonsystematic manner, which lead Sneath and Sokal to the creation of the field of numerical taxonomy, one of the first manifestations of

Simpon's *Systematics* (19,20). Similarly, OTUs were used to determine relationships between traits; however, it allowed for the use of a significantly larger data set.(4,21–23) Sneath and Sokal built an empirical methodology that allowed overall similarity to be distinguished from the phylogenetic relationship by calling it a phenetic relationship, indicating that it be judged by the phenotype of the organism and not its phylogeny.(20) Doing so, created two fields of study, those who subscribed to the idea of evolution's effect on speciation and those that chose to stay firmly within the bounds of phenetics. While numerical phenetics is well accepted in the botanical systems, those who studied cladistics and evolutionary classification vigorously objected to the use of numerical phenetics. (24) Those objecting raise the issue of *homoplasy* (e.g. both octopi and humans have complex compound eyes; but in octopi, they are of a separate derivation than humans).(25) For this reason, those in the field began searching for a more effective and accurate way of classifying organisms, leading to the creation of the Biological Species Concept and The Modern Synthesis.

Biological Species Concept (BSC)

"Species are groups of actually or potentially interbreeding natural populations that are reproductively isolated from other such groups"(26) as defined in the Biological Species Concept introduced by Dobzhanksy in 1937, Muller and Mayr in 1942.(26,27) This concept characterizes two major barriers that allow for speciation to occur: prezygotic and postzygotic. (26)

Prezygotic barriers are centered primarily on the sexual behavior of an organism. Before an organism can begin the process of genetic inheritance, that organism must be in the same geographic vicinity as a potential partner. This simple obstacle has developed into the fields of ecological and geographic speciation introduced by Van Valen(3) and characterized by Barraclough(28), respectively(29). Only thereafter can a pair of organism's ethology come into

play, such as a male's plumage or courtship behavior, which can pleiotropically produce sexual isolation and conceivably selection.(12,29) Sexual selection's effects on speciation were first introduced by Haskins and Haskins in 1949 and further studied by Coyne and Orr in 1998.(12,26,28,30). Logically, after two organisms have interacted and successfully courted each other, the next obstacle would be the physical mechanics involved in the act of copulation, such as female's control of sperm usage, or male-male sperm competition within multiple inseminated females.(12,29) However, the latter of the three examples could be effected by gametic compatibility as well. (26) Nearly all prezygotic isolation mechanisms occur sympatrically, potentially allowing for the formation of several species from a single population. (31–36) However, focusing on the reproductive compatibilities and patterns of interbreeding can cause a misrepresentation of the significance of hybridization among differentiated taxa"(37), which drives the conversation further towards the next of the two barriers responsible for speciation within this concept. (38)

Postzygotic barriers indicatively occur after fertilization has taken place and manifests as either hybrid inviability or sterility, effectively an unfit organism.(12,26,39) "The genetics of hybrid sterility and inviability quickly grow complex as species diverge"(12,40); however, it complicates where the line is to be drawn between one species to another. The Dobzhansky-Muller model sought to resolve some of this complexity through the following central assumption; "Alleles cause no sterility or inviability on their normal "pure species" genetic background".(40) As well, it posits that postzygotic isolation arises in allopatry as a side-effect of ordinary evolutionary divergence(39) and naturally implies that the genetics of speciation will grow very complex very fast. (39) Significantly this shows that the evolution of hybrid sterility or inviability need not involve any intermediate, maladaptive step, later characterized as a

missense mutation. (40) Yet, “doubling of genetic divergence can cause at least a fourfold increase in the expected number of incompatibilities contributing to hybrid sterility, therefore genetic analysis of long-diverged-species might grossly overestimate the number of substitutions needed to obtain speciation”.(39) The difficulty lies in how to methodically characterize species in growing gene pools where the genetic variation within them is consistently rising. The biological species concept is widely used because it describes the present and possible evolutionary future, as well as the concepts of limited gene exchange. However it has limitations as it does not consider asexual species, evolutionary history, and it only focuses on a small window of time, typically that which can be observed in living organisms. (33,34) For those reasons, those in the field have looked for a better species concept that does not have these issues.

Phylogenetic Species Concept (PSC)

The PSC began when Huxley introduced The Modern Synthesis in 1887; joining Mendelian genetics with Darwinian theories on evolution(1). The BSC and PSC did not exist dichotomously; Dobzhansky, an author of the BSC, stated, “genetics has so profound a bearing on the problem of the mechanisms of evolution that any evolution theory which disregards the established genetic principles is faulty at its source”.(2) However, PSC’s popularity did not grow until major advancements in the field of systematics had taken place, allowing for the further development of the study of evolutionary biology. (38) Early studies focused on the development and analysis of cladistics first introduced by Willi Hennig in 1950 and introduced a phylogenetic species as “an irreducible basal cluster of organisms that is diagnosably distinct from other such clusters, and is the smallest monophyletic group of common ancestry”.(13,25,37,41) Cladistics manifests as cladograms or phylogenetics trees where three different relationships can be

displayed: Monophyletic, Paraphyletic, Polyphyletic. (13) Throughout the late 1900s, cladistics was applied and intertwined with the Biological, Ecological, Genealogical, and Evolutionary species concepts, amassing a swath of information to be studied. Leading to Wheeler and Nixon's statement:

> "The militant view that systematists need to embrace is that the responsibility for species concepts lies solely with systematists. If we continue to bow to the study of process over pattern, then our endeavors to elucidate pattern become irrelevant."(42)

Introduction Modern Systematics

The National Institute of Health (NIH) took Wheeler and Nixon's statement to heart and founded both the National Center for Biotechnology Information (NCBI), which worked to create and maintain biomedical databases(43), and the Human Genome Project (HGP), "an international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings".(44) The 1990s was a time for advancement in sequencing of genomes from various species and subsequently which created the need for cataloging of all this new, highly granular, information. Genbank(45) was created primarily to accommodate nucleic acid and amino acid records from data sources around the world namely, EMBL-Bank, (DDBJ) and, NCBI; who make up the International Nucleopeptide Sequence Database (INSDC). (45) As the name implies, GenBank is a repository of genetic information, but it does provide a clean, curated dataset for researchers to use, for this reason, Refseq was founded. (45) The RefSeq project leverages data submitted to the (INSDC) against a combination of computation, manual curation, and collaboration to produce a standard set of stable,

non-redundant reference sequences. The database currently represents sequences from more than 55,000 organisms, ranging from a single record to complete genomes. (46)

The GENE and the GENOME project are products of the INSDC and hosted by the NCBI. GENE creates analytical support for this incoming data; while the former worked to curate accommodate the contextual needs of each genetic records (i.e. map, sequence, expression, structure, function, citation, and homology data).(47) GENOME developed BLAST, a very powerful alignment tool; As well it is the home of the Human Genome Project as well as many other sequencing/annotations projects.(44,48) The United States was not the only country leading this effort; Japan's KEGG created an integrated database resource consisting of fifteen manually curated databases and a computationally generated database in four categories which also contains GENOME and GENES, this integration made possible through the DBBJ membership in INSDC.(49) KEGG created the KO database, a collection of genes and proteins, which are curated into intricate pathway maps, that allow the user to visualize their protein of interest within the various pathways in which it is involved and their relationship to the proteins lying upstream and downstream.(49) UniProtKB a member of EMBL-EBI, concerned primarily with the curation of the human genome and related data created partnerships with Swiss's SERI and SIB, together, they were the first to publish the complete human proteome in 2008. (50–53) There are numerous other organizations that have made advancements in the field of data collection and analysis, however, for the purposes of this study, we will limit our review to those mentioned above and utilized in our work.

Applying Bioinformatics to the species problem

In the case of translational medicine and drug discovery, it is imperative to have a thorough understanding of the genetics at work within the species of interest in order to understand its pharmacodynamics and how those dynamics will translate to humans. However, due to the issues within the aforementioned methodologies for species classification, it becomes difficult to predict the exact effect a treatment will have in a species of interest and if that treatment will subsequently translate to humans. For this reason, there is a need for a framework for finding homologous pairs that: perform the same function, are evolutionarily stable and are in a similar physiological location within the taxonomic class. Here we present such a framework, which leverages relevant fundamental genetic data from the INSDC, DDBJ, and EMBL databases in a format that can be connected to the human and animal health information, in order to provide both contextual and computational base for which translational research to be conducted.

Chapter 2 - Methods and Results

Evolutionary Rate Framework

The following methods are adapted from the Wisconsin package, a software developed originally at the University of Wisconsin-Madison, and sold through several corporate entities since. Their method was adapted from Louis, who adapted the computational methods from Li et al, who implemented Kimura's Two Parameter method. (54–56)

The script begins by searching GenBank's directories for all nucleic and amino acid pairs (325,952 pairs) in mammals (a.k.a gbmam). It then performs reverse translation, using the provided amino acid sequence, in order to verify that the provided nucleic acid (NA) sequence codes for the provided amino acid (AA) sequence; if extra NA sequence is provided on either end, the excess is removed, leaving only the coding sequence (CDS). Due to the large number of incomplete or partial records in GenBank, this leaves approximately 93,000 pairs. Each of the 93,000 NA sequences is aligned to one another using a global alignment algorithm (Needleman–Wunsch**)** provided by BioPython(57)**;** creating approximately 8,605,000,000 alignments.

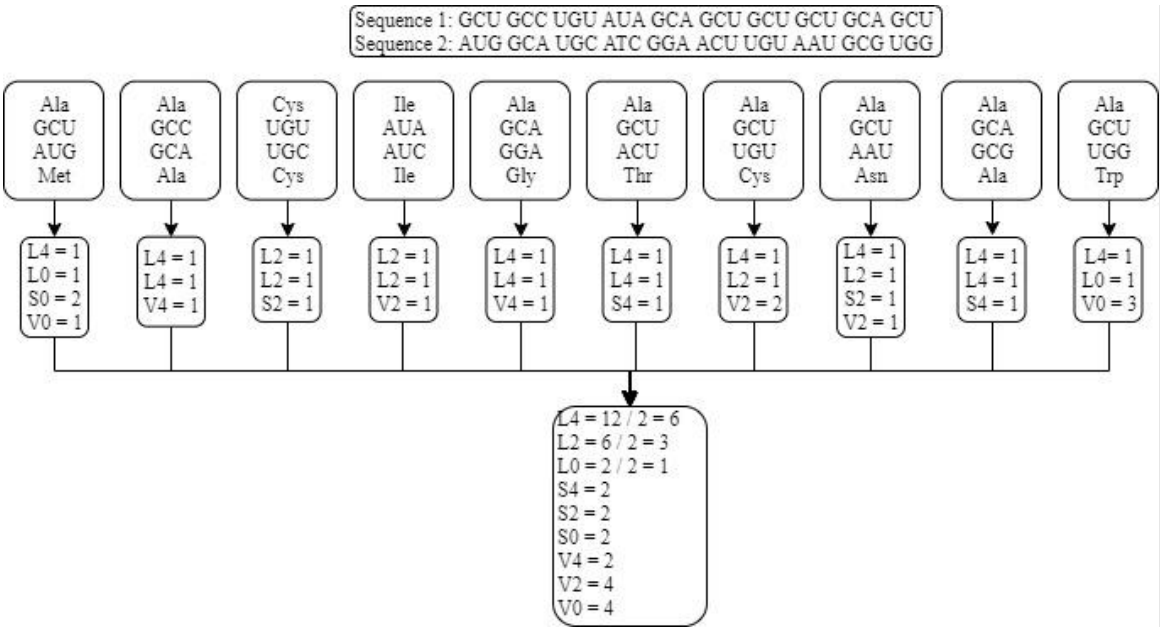


Figure 1. Workflow for calculation of Codon Degeneracy and Tranversional/Transitional Frequency

Depicted in Figure 1, two sequences are aligned and broken up into codon pairs. Each codon in a set is classified into one of three codon degencery categories (L_0, L_2, L_4) using the key in Figure 2.

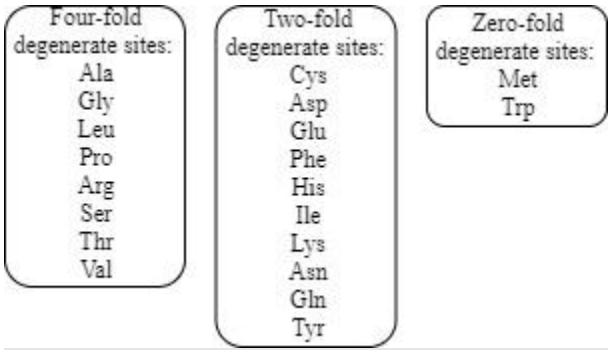


Figure 2. Key for Codon Degeneracy

The average codon degeneracy for the alignment is obtained by counting the number of non-degenerate (L_0), twofold degenerate (L_2), and fourfold degenerate (L_4) amino acids and dividing the sum of each by two as depicted in Figure 1. Each codon within a set is compared to the other and the frequency of transversion (V_i) which is changed from

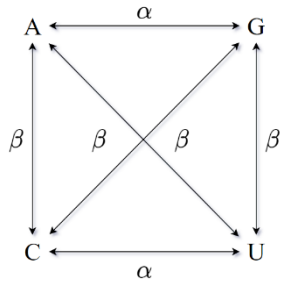
(A<->C, A<->U, U<->G, C<->G) and transition (S_i) (A<->G, C<->U) are collected and sorted by their codon degeneracy designation ($S_0, S_2, S_4, V_0, V_2, V_4$).

$$P_0 = \frac{S_0}{L_0} \quad P_2 = \frac{S_2}{L_2} \quad P_4 = \frac{S_4}{L_4}$$

$$Q_0 = \frac{V_0}{L_0} \quad Q_2 = \frac{V_2}{L_2} \quad Q_4 = \frac{V_4}{L_4}$$

Figure 3. Transitional and Trasnsversional distance equations

Transitional (P_i) and Tranversional (Q_i) distances are calculated by taking the ratio of the frequency of each transitional or transversion total and each codon degeneracy total respective of its codon degeneracy designation shown in Equations above.



$$a_0 = 1 - (2 * (P_0 - Q_0) \qquad b_0 = 1 - (2 * (Q_0)$$
$$a_2 = 1 - (2 * (P_2 - Q_2) \qquad b_2 = 1 - (2 * (Q_2)$$
$$a_4 = 1 - (2 * (P_4 - Q_4) \qquad b_4 = 1 - (2 * (Q_4)$$
$$\alpha_0 = (-0.5 * \ln(a_0)) + (0.25 * \ln(b_0)) \qquad \beta_0 = -0.5 * \ln(b_0)$$
$$\alpha_2 = (-0.5 * \ln(a_2)) + (0.25 * \ln(b_2)) \qquad \beta_2 = -0.5 * \ln(b_2)$$
$$\alpha_4 = (-0.5 * \ln(a_4)) + (0.25 * \ln(b_4)) \qquad \beta_4 = -0.5 * \ln(b_4)$$

Figure 4 Kimura's Two Parameter Test - Left: Depiction of Transversion and Transition and its application to Kimura's Two Parameter test. Right: Depiction of calculations used in Kimura's Two Parameter Test.

The Transitional (L_i) and tranversional (V_i) distances are used to calculate ($a_{0,}, a_{2,}, a_{4,}, b_{0,}, b_{2,}, b_{4,}$), which are then used in Kimura's two-parameter test ($\alpha_{0,}, \alpha_{2,}, \alpha_{4,}, \beta_{0,}, \beta_{2,}, \beta_{4,}$) which estimates the true number of transitional and tranversional substitutions for each degencery designation per site, as depicted in Figure 4. Ultimately the ratio of non-synonymous change ($K_{a,}$) and synonymous change ($K_{s,}$) are calculated using the codon

degencery frequency values ($L_0, L_{2,} L_4$) and the values generated from Kimura's two-parameter test ($\alpha_{0,}, \alpha_{2,}, \alpha_{4,}, \beta_{0,}, \beta_{2,}, \beta_{4,}$), depicted in Figure 5.

$$K_s = \frac{\beta_{4,} + \left(L_{2,} * \alpha_{2,}\right) + (L_4 * \alpha_{4,})}{L_{2,} + L_{4,}} \qquad K_a = \frac{\alpha_0 + (L_0 * \beta_0) + (L_2 * \beta_{2,})}{L_{0,} + L_2}$$

$$\frac{K_a}{K_s}$$

Figure 5. Top: K_a, K_s calculation. Bottom: K_a, K_s Ratio

However, not all sequence pairs will reach this stage, if ($a_{i,}$ *or* $b_{i,}$) are not greater than zero, then the respective α_0 or β_0 will be undefined and subsequently, the calculation will end at this stage. Similarly, if the sum of (L_2 and L_4) or (L_2 and L_2) equal zero then K_a or K_s will be undefined respectively, the same logic applies if $K_s$is equal to zero. For these reasons, the result of the script will have less than 8,605,000,000 results.

The results generated by the script are classified into three categories based upon the calculated $\frac{K_a}{K_s}$ value. Gene pairs that have a $\frac{K_a}{K_s}$ less than 1 are considered to be under negative or purifying selection, those genes with $\frac{K_a}{K_s}$ greater than 1 are under positive or Darwinian selection and those with a $\frac{K_a}{K_s}$ equal to 1 are under neutral selection. The $\frac{K_a}{K_s}$ values were calculated regardless of the species of origin, this was done in order to find mislabeled genes within GenBank gbmam dataset. For this reason, it was imperative to align gene pairs with those in Refseq in order to determine if those genes are comparable, meaning they are the same or similar genes.

Ontological Framework

Sequence pairs that have $\frac{K_a}{K_s}$ values are aligned with sequences from Refseq, again, using the Needelman-Wunsch alignment method. This creates a list of the species for which each gene exists. This list of species can be used to create a gene tree, which will depict if a gene exists in multiple species. If a gene is shared across multiple species, their respective amino acid sequences can be used to search UniprotKB, which serves as a connection to the KEGG database by providing the KEGG ID. If a KEGG ID exists, that ID is then used to search the KEGG database to determine if the gene pairs are found to be in the same cellular milieu and pathway. The script then creates a record containing: Unique comparison ID (Internal Use), Genbank ID, RefSeq ID, Uniprot ID, KEGG ID, and the sequences.

Development in an Open Source Environment:

The aforementioned framework is implemented on a server that runs Ubuntu 16.04 LTS server operating system. Python 3.0 was chosen due to its vast community support and its ability to multi-thread and be run in parallel, which improves the speed at which tasks are completed, and is well supported in Ubuntu Linux environment. The database system chosen was MySQL for its ability to quickly store and retrieve structured information and it is well supported in the Ubuntu Linux system.

Chapter 3 - Discussion

Translational Medicine

The ontological framework's objective is to provide a curated database of mammalian genes with information pertaining to their evolutionary stability and conservation across other mammals. A gene's evolutionary stability is a strong indicator of its conservation either within or across species. If a gene pair is not under selection and is conserved; it can be inferred that the gene pair could be a good target for pharmaceutical development. However, highly conserved genes may be highly expressed in an organism, and for that reason, designing a highly specific pharmaceutical could be challenging, as the gene may be doing various functions in various tissues, which would result in various pharmaceutical effects that in some cases would be adverse to the species in question. For those in the field of translational medicine, challenges such as these are an everyday occurrence. Many pharmaceuticals have failed due to these challenges in the understanding of the translational aspect of a pharmaceutical's development. This failure commonly manifests, when the drug is taken from a "successful" animal trial and implemented into human testing. Where, it may either "fail early" meaning that at Phase 1 the drug was causing serious side effects, or it will fail in late Phase 2/early Phase 3 due to lack of efficacy. This failure occurs because the intended receptor the pharmaceutical is targeting in the human may not be in the same cellular milieu or possibly not in the same pathway, as the tested animal model. This gap in knowledge has affectionately been named the "Valley of Death". However, pharmaceutical companies work hard to ensure that their pharmaceuticals do not fail; an early Phase 1 failure can cost upwards of 100 million dollars and late Phase 2/ early phase 3 can be upwards of 200-300 million dollars. This immense cost is a risk to any company in the field and for that reason, pharmaceutical companies have shifted away from doing the pre-phase

2 research and development, in favor of buying smaller biotech companies who have already successfully completed those steps. The immense cost to pharmaceutical companies has created an incentive in the field of translational research, in that, funding is available for those innovating the field as the pharmaceutical industry attempts to mitigate these costly failures.

Translational research should go both ways; when a drug is developed for a human, an animal model is used, therefore, at the end of the drug development pipeline, two drugs should exist: One for the human and one for the animal, used as the animal model. The pharmacokinetics, as well as the efficacy and safety data collected during the animal trial, can then be used to design drugs for other mammals with similar cellular targets and systemic physiology. Designing drugs with multiple species in mind, can mitigate the risk of financial investment as it is diversified across multiple animals and subsequently multiple revenue streams.

Integration of human and animal health data

There has been a recent rise in intra/inter-health record analysis due to advancements in electronic medical record (EMR) software(s) and their ability to provide diagnostic assistance tools to physicians. This advancement has allowed hospital systems, as a whole, to gain insight into their patient population; allowing them to optimize patient care, reduce short-term re-admittance, and allocate resources proactively to growing departments. The integration of health records to this ontological framework could provide further insight for future pharmaceutical development and reinforce that all side effects are population dependent, and the first step to their mitigation is identification and classification of their cause. The collection for side effects stratified based upon hospital or region could provide better insight into specific pharmaceuticals within those populations and subsequently, expand our understanding of the genetic diversity

within and across human populations. However, the use of health records to assist phase 4 drug monitoring has yet to be implemented on a large scale partly because the process of anonymizing human health records and sharing that data with other organizations is complex. Further, the present method for tracking a drug's performance, in both humans and animals, is by looking at the FDA adverse drug events (ADE) database(58). There you will find millions of ADEs, however, there is not a standard method for discerning if a reported symptom is caused by the reported drug or if it is a symptom that patient is suffering independent of the reported treatment or other coexisting conditions such as concomitant disease, diet or multiple drugs being administered. (59)

Human healthcare institutions are a large source of human health data, and a vast amount of time and resources were implemented to obtain and curate that data in a meaningful format. However, these institutions are not owners of this information; as it is the property of the patient. For this reason, there has been a push to enroll patients into data-sharing agreements, which allow health care institutions to share human health data through organizations like iShare medical,(60) which allows patients to share their own medical records. Research programs like 1Data(61) at Kansas State University(62) could then combine that human data with animal health information along with ADE and genetics information to build out a more robust translational research framework.(59,63)

There is an additional layer of translation that can occur within just animal species. This includes developing drugs for multiple breeds for example of dogs, cats or cattle. In addition, disease manifestations may be very different between different animal species and humans, or even within breeds of a species. Classic examples, many having a pharmacogenomic basis, include species and breed differences in nutritional requirements due to very different

gastrointestinal physiology, drug-metabolizing enzymes (cytochrome p450 system polymorphisms) and drug transporter systems (e.g. P-glycoprotein regulated by the MDR1 or ABCB1 gene). The later results in a greatly increased sensitivity of collie dogs to ivermectin exposure compared to most other breeds. In addition to these pharmacogenomic variables, drug development across species is also dependent on different body sizes (e.g. a mouse to a horse) which results in the need for allometric scaling or using detailed blood-flow sensitive pharmacokinetic models. Great progress has been made in these areas, however, successful translation is a multifactorial and complex situation that requires parallel studies in comparative genomics as well as pharmacokinetic and cellular response systems. (64)

Evolutionary analysis

It is understood that K_a, K_s, and $\frac{K_a}{K_s}$ constitute the amino acid substitution rate, mutation rate, and selective pressure, respectively. However, using $\frac{K_a}{K_s}$ or K_s essentially ignores codon bias; therefore, while we consider K_s to be the mutation rate it's better characterized as the fixation rate for synonymous mutations. (65) It also understood that sequences that show high codon bias is associated with lower K_s values. (66) It then comes as no surprise that high K_s values are correlated with high $\frac{K_a}{K_s}$ values because genes that are under heavy codon bias constraint are likely under even higher nonsynonymous mutation constraint. Although K_a, K_s, and their ratio have some translational specificity, it is prudent to consider all three when characterizing the influence of their conversation and how that affects two gene's relationship to one another as well as their ability to be targeted for drug development. "Our understanding of intragenomic mutation rate variation remains limited and is drawn from a relatively small number of model organisms."(11) For this reason, we sought to implement a framework that would alleviate the

need for a researcher to consult multiple data sources in order for them to conduct translational genomic analysis; thereby accelerating the field's understanding of intragenomic mutation rates by providing an ontology for which they can cross-reference their findings. In order to extend this analysis to the creation of phylogenic trees based on genetic similarity, we would then need to calculate a 'total species divergence' based on the composite of the total pathway divergences. This would allow for clustering analysis to take place while maintaining the integrity of the research aim to find good model organisms for drug target pathways. In order to alleviate codon bias, and to better serve cluster analytical methods it may be beneficial to implement Grantham distance and BLOSUM62 in order to interpret instances where K_a is equal to K_s and where K_s is far grater than K_a.

Absolute mutation rates (i.e. the number of mutations per site and generation) are nonuniform across the genome, and the implementation of genome-wide mutation rate variation should be taken into consideration in order to interpret the genomic landscape accurately. (67,68)

Application to the Species problem

It is imperative to subscribe to a system of speciation that allows for the accurate and consistent investigation of organisms; “Understanding the demographic and evolutionary history of population and species pairs is necessary to generate expected patterns of genomic differentiation.”(11) Through further development in the understanding of humans and animals and as their genetic and evolutionary history become more clear, it may become easier to classify species and be able to develop more stringent and clear criteria on what separates one species from another.

Further Development of the Framework

The next two steps in developing this framework for end-users is viability and speed. At this time, the developed framework is not readily available to the public. It is a collection of command-line tools that do not have a graphical user interface (GUI) nor are the individual tools user-friendly. In order to mitigate this, the construction of a web-based front-end would be beneficial, as it would provide a venue for end-users to access and use it without having to learn Python(69) or any other coding language. However, when applications are web-based, precautions must be put into place in order to protect the integrity of the app as well as the security of any data stored within it. For that reason, Django (70), web-framework written in Python as well, would be a logical choice as it is known for its security and flexibility in developing a web-based interface. Django would allow for several levels of user access (i.e. Database administrators, end-users) as well as handling user-authentication and management, saving time in development and technical support. Django's (70) framework can be used to allow researchers to submit their data, and view it among the 1Data (61) dataset as well as access and use analytical tools such as $\frac{K_a}{K_s}$ calculation or the ontological framework. (63) As well, the Django web framework would allow for the future implementation of other analytical tools.

Speed is the next challenge, as stated above in the $\frac{K_a}{K_s}$ calculation there are over 8 billion computations to complete. At present, the server that performs this task can complete up to 20 calculations per second (20 cores at one calculation per core per second) which would take approximately 13 years to complete. In order to mitigate this several steps could be taken: the simplest would be upgrading the server from Ubuntu version 16 to Ubuntu version 18, which would increase the scripts run time. However, it is likely that the completion time would still be measured in years. The next step might be moving all data stored in MySQL to PostgreSQL,

which could take months off the completion time, as PostgreSQL indexing algorithm is much simpler and more robust. However, the most important step would be, optimization of the python scripts and potentially changing the algorithm to allow easier parallelization, to allow it to be implemented on Beocat which has more than 1000 cores. This would allow the script to be completed in as few as 14 weeks.

Chapter 4 - Conclusion

Researchers in all fields of biology rely on a clear concise definition of species, whether it is using the methods defined above or using methods yet to be written; it is paramount for the further development of the life sciences that a reproducible and abstractable definition be determined and implemented. The phylogenetic species concept provides much of the needed framework for speciation determination; however, it lacks a computational component to discern between two closely related species, and for this reason, requires further development.

References

1. Huxley J. Evolution The Modern Synthesis. 1st ed. New York & London: Harper & Brothers; 1942. 645 p.

2. Dobzhansky T. Genetics and The Origin of Species. New York: Morningside Heights: Columbia University Press; 1941.

3. Coyne JA. Ernst Mayr and the Origin of Species. Evolution. 1994;48(1):19–30.

4. McNeill J. PHENETIC CLASSIFICATION SYSTEMS [Internet]. Powerpoint presented at: Plant Taxonomy (BIOL308); 2007 [cited 2019 Jul 17]; College of St. Benedict/St. John's University. Available from: https://employees.csbsju.edu/SSAUPE/biol308/Lecture/Classification/phenetic_class.htm

5. History of herbalism. In: Wikipedia [Internet]. 2019 [cited 2019 Jul 18]. Available from: https://en.wikipedia.org/w/index.php?title=History_of_herbalism&oldid=904427920

6. Aristotle's theory of universals. In: Wikipedia [Internet]. 2018 [cited 2019 Aug 9]. Available from: https://en.wikipedia.org/w/index.php?title=Aristotle%27s_theory_of_universals&oldid=855352191

7. Wyckoff, Gerald J. L Lee, Rachael Allen. Evolution: Principles and Practice. Powerpoint presented at: Introduction to Evolution; 2017; University of Missouri Kansas City.

8. History of Evolution | Internet Encyclopedia of Philosophy [Internet]. [cited 2019 Oct 20]. Available from: https://www.iep.utm.edu/evolutio/

9. Epicurus | Internet Encyclopedia of Philosophy [Internet]. [cited 2019 Oct 20]. Available from: https://www.iep.utm.edu/epicur/

10. Witt NWD. Epicurus and His Philosophy. U of Minnesota Press; 1954. 398 p.

11. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, et al. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. Journal of Evolutionary Biology. 2017;30(8):1450–77.

12. Coyne JA, Orr HA. The evolutionary genetics of speciation. Philos Trans R Soc Lond B Biol Sci. 1998;353(1366):287–305.

13. Geonyzi A. Phenetic, cladistic, cladogram, phylogenetics [Internet]. Powerpoint presented at; 2017 Mar 20 [cited 2019 Jul 16]; Online. Available from: https://www.slideshare.net/geonyzl/phenetic-cladistic-cladogram-phylogenetics-73362398

14. Wyckoff, Gerald J. L Lee, Rachael Allen. Brief History of Evolution. Powerpoint presented at: Introduction to Evolution; 2017; University of Missouri Kansas City.

15. *Genera Plantarum*. In: Wikipedia [Internet]. 2019 [cited 2019 Jul 18]. Available from: https://en.wikipedia.org/w/index.php?title=Genera_Plantarum&oldid=888326773

16. Antoine Laurent de Jussieu. In: Wikipedia [Internet]. 2018 [cited 2019 Jul 18]. Available from: https://en.wikipedia.org/w/index.php?title=Antoine_Laurent_de_Jussieu&oldid=835522748

17. *Prodromus Systematis Naturalis Regni Vegetabilis*. In: Wikipedia [Internet]. 2019 [cited 2019 Jul 18]. Available from: https://en.wikipedia.org/w/index.php?title=Prodromus_Systematis_Naturalis_Regni_Vegetabilis&oldid=901864273

18. Augustin Pyramus de Candolle. In: Wikipedia [Internet]. 2019 [cited 2019 Jul 18]. Available from: https://en.wikipedia.org/w/index.php?title=Augustin_Pyramus_de_Candolle&oldid=906650631

19. Simpson GG. Principles of animal taxonomy. 2nd ed. New York: Columbia University Press; 1961. 6–7 p.

20. Sneath PHA, Sokal RR. Numerical Taxonomy: The principles and practice of numerical classification. Kennedy D, Park RB, editors. San Francisco: W. H. Freeman and Company; 1973. (A series of books in biology).

21. Faiza M. What is Numerical Taxonomy? How is it useful? [Internet]. Bioinformatics Review. 2016 [cited 2019 Jul 18]. Available from: https://bioinformaticsreview.com/20160225/what-is-numerical-taxonomy-how-it-works/

22. Numerical Taxonomy an overview | ScienceDirect Topics [Internet]. [cited 2019 Jul 18]. Available from: https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/numerical-taxonomy

23. Raina V, Nayak T, Ray L, Kumari K, Suar M. Chapter 9 - A Polyphasic Taxonomic Approach for Designation and Description of Novel Microbial Species. In: Das S, Dash HR, editors. Microbial Diversity in the Genomic Era [Internet]. Academic Press; 2019 [cited 2019 Jul 18]. p. 137–52. Available from: http://www.sciencedirect.com/science/article/pii/B9780128148495000095

24. Duncan T, Baum BR. Numerical Phenetics: Its Uses in Botanical Systematics. Annual Review of Ecology and Systematics. 1981;12:387–404.

25. Wyckoff, Gerald J. L Lee, Rachael Allen. Phylogenetic Concepts. Powerpoint presented at: Introduction to Evolution; 2017; University of Missouri Kansas City.

26. Wyckoff, Gerald J. L Lee, Rachael Allen. What are Species? Powerpoint presented at: Introduction to Evolution; 2017; University of Missouri Kansas City.

27. Shapiro BJ, Leducq J-B, Mallet J. What Is Speciation? PLoS Genet [Internet]. 2016 Mar 31 [cited 2019 Jul 16];12(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4816541/

28. Barraclough TG, Vogler AP. Detecting the Geographical Pattern of Speciation from Species-Level Phylogenies. The American Naturalist. 2000 Apr 1;155(4):419–34.

29. Rice WR. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature. 1996 May 1;381(6579):232–4.

30. Haskins CP, Haskins EF. The Role of Sexual Selection as an Isolating Mechanism in Three Species of Poeciliid Fishes. Evolution. 1949;3(2):160–9.

31. Kondrashov AS, Kondrashov FA. Interactions among quantitative traits in the course of sympatric speciation. Nature. 1999 Jul;400(6742):351.

32. Mina MV, Mironovsky AN, Dgebuadze Y. Lake Tana large barbs: phenetics, growth and diversification. Journal of Fish Biology. 1996;48(3):383–404.

33. Berrebi P, Valiushok D. Genetic divergence among morphotypes of Lake Tana (Ethiopia) barbs. Biological Journal of the Linnean Society. 1998;64(3):369–84.

34. Schliewen UK, Tautz D, Pääbo S. Sympatric speciation suggested by monophyly of crater lake cichlids. Nature. 1994 Apr;368(6472):629–32.

35. Johnson TC, Scholz CA, Talbot MR, Kelts K, Ricketts RD, Ngobi G, et al. Late Pleistocene Desiccation of Lake Victoria and Rapid Evolution of Cichlid Fishes. Science. 1996 Aug 23;273(5278):1091.

36. Berlocher SH, Howard DJ. Endless forms : species and speciation [Internet]. New York City: Oxford University Press; 1998. Available from: http://www.worldcat.org/oclc/37545522

37. Zink RM, McKitrick MC. The Debate over Species Concepts and Its Implications for Ornithology. The Auk. 1995 Jul 1;112(3):701–19.

38. Avise JC, Wollenberg K. Phylogenetics and the origin of species. PNAS. 1997 Jul 22;94(15):7748–55.

39. Orr AH, Turelli M. The evolution of postzygotic isolation: accumulating dobzhansky-muller incompatibilities. Evolution. 2001;55(6):1085–94.

40. Orr HA. The Population Genetics of Speciation: The Evolution of Hybrid Incompatibilities. Genetics. 1995 Apr;139(4):1805–13.

41. Hennig W. Phylogenetic Systematics. Stuttgart, Germany: Stoatliches Mueseam; 1965. 20 p.

42. Wheeler QD, Nixon KC. Another Way of Looking at the Species Problem: A Reply to De Queiroz and Donoghue. Cladistics. 1990;6(1):77–81.

43. National Center for Biotechnology Information (NCBI) [Internet]. National Center for Biotechnology Information. 1988 [cited 2019 Sep 15]. Available from: https://www.ncbi.nlm.nih.gov/home/about/mission/

44. What is the Human Genome Project? [Internet]. Genome.gov. [cited 2019 Sep 15]. Available from: https://www.genome.gov/human-genome-project/What

45. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013 Jan;41(Database issue):D36–42.

46. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(Database issue):D733–45.

47. National Center for Biotechnology Information (NCBI) [Internet]. [cited 2019 Sep 16]. Available from: https://www.ncbi.nlm.nih.gov/gene

48. Home - Genome - NCBI [Internet]. [cited 2019 Sep 16]. Available from: https://www.ncbi.nlm.nih.gov/genome

49. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan 4;45(Database issue):D353–61.

50. Breuza L, Poux S, Estreicher A, Famiglietti ML, Magrane M, Tognolli M, et al. The UniProtKB guide to the human proteome. Database (Oxford) [Internet]. 2016 Feb 19 [cited 2019 Jun 26];2016. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4761109/

51. The European Bioinformatics Institute < EMBL-EBI [Internet]. [cited 2019 Sep 16]. Available from: https://www.ebi.ac.uk/

52. Flegel S. What we do [Internet]. [cited 2019 Sep 16]. Available from: https://www.sib.swiss/about-sib/what-we-do

53. SERI SS for E Research and Innovation. State Secretariat for Education, Research and Innovation [Internet]. 2019 [cited 2019 Sep 16]. Available from: https://www.sbfi.admin.ch/sbfi/en/home/das-sbfi/das-sbfi.html

54. Li C-W, Jheng B-R, Chen B-S. Investigating genetic-and-epigenetic networks, and the cellular mechanisms occurring in Epstein–Barr virus-infected human B lymphocytes via big data mining and genome-wide two-sided NGS data identification. PLOS ONE. 2018 Aug 22;13(8):e0202537.

55. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980 Dec;16(2):111–20.

56. University of Wisconsin-Madison. Diverge [Internet]. 2002 [cited 2019 Sep 30]. Available from: http://www.biology.wustl.edu/gcg/diverge.html#algorithm

57. Biopython: freely available Python tools for computational molecular biology and bioinformatics | Bioinformatics | Oxford Academic [Internet]. 2019 [cited 2019 Oct 8]. Available from: https://academic.oup.com/bioinformatics/article/25/11/1422/330687

58. openFDA [Internet]. [cited 2019 Oct 7]. Available from: https://open.fda.gov/

59. Xu X, Mazloom R, Goligerdian A, Staley J, Amini M, Wyckoff G, et al. Making Sense of Pharmacovigilance and Drug Adverse Event Reporting: Comparative Similarity Association Analysis Using AI Machine Learning Algorithms in Dogs and Cats. Topics in Companion Animal Medicine. 2019 Sep 30;100366.

60. iShare Medical [Internet]. iShare Medical. [cited 2019 Oct 7]. Available from: https://www.isharemedical.com/

61. 1Data: Improving the lives of humans and animals [Internet]. [cited 2019 Sep 18]. Available from: https://olathe.k-state.edu/research/centers-institutes/1data/

62. Kansas State University [Internet]. [cited 2019 Oct 7]. Available from: https://www.k-state.edu/

63. Staley J, Mazloom R, Lowe P, Newsum C, Jaberi M, Riviere J, et al. Novel data sharing agreement to accelerate big data translational research projects in the one health sphere. Topics in Companion Animal Medicine. 2019 Oct 1;100367.

64. Riviere JE, Papich MG, editors. Veterinary Pharmacology and Therapeutics. 10th ed. Wiley-Blackwell; 2018. 1552 p.

65. Sharp PM, Li WH. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Molecular Biology and Evolution. 1987 May 1;4(3):222–30.

66. Wyckoff GJ, Malcom CM, Vallender EJ, Lahn BT. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate: Trends in Genetics. TRENDS in Genetics. 2005 Jul;21(7):381–5.

67. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nature Reviews Genetics. 2011 Nov 1;12(11):756–66.

68. Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD. Extensive de novo mutation rate variation between individuals and across the genome of Chlamydomonas reinhardtii. Genome Res. 2015 Nov;25(11):1739–49.

69. Welcome to Python.org [Internet]. Python.org. [cited 2019 Oct 20]. Available from: https://www.python.org/

70. The Web framework for perfectionists with deadlines | Django [Internet]. [cited 2019 Oct 20]. Available from: https://www.djangoproject.com/