Regis University
ePublications at Regis University

All Regis University Theses

Winter 2019

The Future of Biomath: Growing Beyond Collaboration

Kyle Weishaar
Regis University

Follow this and additional works at: https://epublications.regis.edu/theses

Recommended Citation

Weishaar, Kyle, "The Future of Biomath: Growing Beyond Collaboration" (2019). *All Regis University Theses*. 937.
https://epublications.regis.edu/theses/937

This Thesis - Open Access is brought to you for free and open access by ePublications at Regis University. It has been accepted for inclusion in All Regis University Theses by an authorized administrator of ePublications at Regis University. For more information, please contact epublications@regis.edu.

THE FUTURE OF BIOMATH: GROWING BEYOND COLLABORATION

A thesis submitted to

Regis College

The Honors Program

in partial fulfillment of the requirements

for Graduation with Honors

by

Kyle Weishaar

December 2019

Thesis written by

Kyle Weishaar

Approved by

Thesis Advisor

Thesis Reader

Accepted by

Director, University Honors Program

Contents

acknowledgments

Thank you to Dr. Tim Trenary for advising this thesis and Dr. Max Boeck for reading this thesis.

Chapter 1

Introduction

Biomathematics as a field has grown substantially over the last 50 years. It has found success in modeling biological phenomena in a variety of areas ranging from ecology to molecular biology [Mackey and Maini, 2015]. Furthermore, the continued development of biomath may be invaluable in understanding current challenges in biology, such as predicting the effects of climate change on different ecosystems. All successful interdisciplinary research depends different types of scientists having the ability to understand and collaborate well with each other. Traditionally, mathematicians are exclusively trained in theoretical systems, while biologists usually work in experimentally driven laboratory settings. As a result, collaboration can lead to miscommunications and fundamental misunderstandings about both the system being studied and the mathematical tools being used. I argue that until biomath becomes fully integrated into biology such miscommunications cannot be avoided and the field will not reach its full potential.

Chapter 2

Rashevsky's New Biologists

Nicolas Rashevsky is a somewhat controversial figure in mathematical biology. As one of the field's founders, Rashevsky established the first degree granting program for mathematical biology, the first mathematical biology journal, and helped to legitimize the field [Shmailov, 2016, Abraham, 2004]. However, many biologists considered Rashevsky's work to be misguided and generally uninformative. In fact, most of Rashevsky's greatest failures happened because he was unable to convince biologists that his work was relevant [Shmailov, 2016]. Even Rashevsky's legacy is tarnished as many modern day biomathematicians either do not know who he is or distance themselves from his work. [Abraham, 2004]. However, I believe Rashevsky's career is worth studying because the origins of biomath are crucial in understanding the current identity and limitations of the field.

Rashevsky was born on September 20, 1899 and he grew up in a small Ukraine village. In 1919, he earned a PhD in mathematical physics from the University of Kiev. After the USSR invaded the Ukraine, Rashevsky found his career stifled and moved to Prague. His family would eventually immigrate to the United States, where he would work as a researcher and instructor at various institutions. Rashevsky's interest in biology was inspired by a chance meeting with a biologist at a social event

[Shmailov, 2016]. According to one of Rashevsky's students, "[Rashevsky] asked the biologist whether the thermodynamic mechanism on which he was working was the way biological cells divided. He was told that (1) nobody knew how biological cells divided and moreover, (2) nobody could know how biological cells divided, because this was biology" [Rosen, 1991]. Rashevsky was challenged by this notation and his work began to focus on the biological applications of physics.

Rashevsky envisioned mathematical biology as a field that would function similarly as mathematical physics. His version of a mathematical biologist would not be driven by data or observations. Instead, they will adopt the mathematical approach of making fundamental assumptions about a biological system and considering the consequences. He would begin by studying simple cases for which he would eventually add complexity to better reflect reality. Rashevsky argued that this approach will help to overcome the complexity of biology and describe the general characteristics of a system. Rashevsky strongly felt that the theory developed by mathematical biology has value in itself, even if it did not explain or guide experiments [Shmailov, 2016]. However, later in his career he did argue that mathematical biologists should still attempt to guide experiments and should generally strive to make useful predictions about biological systems [Cull, 2007].

For the most part, the biological community did not accept Rashevsky's methods, even going as far as to not credit him with accurate findings. For example, Rashevsky proposed the following model nerve of excitation

$$\frac{de}{dt} = KI - k(e - e_0)$$
$$\frac{di}{dt} = MI - m(i - i_0).$$

The value K, k, M and m are constants, I represents current, e the excitatory process, and i the inhibitory one. Rashevsky verified his model with previously available

experimental results. Around 3 years later Archibald Hill, a physiologist, published a similar model

$$\frac{de}{dt} = KI - k(e - e_0)$$
$$\frac{di}{dt} = M(e - e_0) - m(i - i_0).$$

Despite both models having the same general behavior [Shmailov, 2016], the physiological community credited Hill with the discovery as his method was based on experimentation.

Perhaps, Rashevsky's own habits caused his rejection from the physiological community. In his works, he would often compare himself to great scientists such as Kepler, Newton, and Einstein [Shmailov, 2016]. Worst still, many physiologists felt that Rashevsky made exaggerated claims and was over confident in his conclusions. The most devout experimentalists were also uneasy with his approach of modeling overly simplified cases then gradually introducing complexity. This can partially be explained by Rashevsky not being formally trained in biology [Abraham, 2004]. However, he was not interested in joining the biological community. Rather, he wanted to invent a new field and as a result, he felt the need to distance himself from existing fields within biology.

Rashevsky spend a large portion of his career as a professor at the University of Chicago. However, his inability to connect with experimentalists lead to several cases of internal friction. For example, he was removed from the department of physiology because the department head was an adamant experimentalists and disliked Rashevsky's work. Surprisingly, this ultimately lead to Rashevsky receiving his own research group [Abraham, 2004].

The group's creation represents a major turning point in Rashevsky's career as it enabled him to recruit students, host seminars, and raise the profile of his field

worldwide. However, the group had difficulties getting journals to publish their work. Many thought that their research was too mathematical for biology journals and too biological for mathematical journals[Shmailov, 2016]. This prompter Rashevsky to create the first mathematical biology journal: *The Bulletin of Mathematical Biology*. The journal became the group's main publication outlet and it still exists today. Over time, the group continued to grow and eventually became the Committee on Mathematical Biology, an entity with the power to grant PhDs [Cull, 2007].

Rashevsky's person work focused over a broad set of biological systems but he rarely explored a topic in depth. Over his career, he researched cell biology, nervous systems, and topics in sociology. As a result, his work was often criticized for sharing similarity in methods instead of a common subject matter. Though unusual, this habit is understandable since Rashevsky's main goal was not to be a mathematical biologist. Rather, he was attempting to pioneer a new field and to his credit, many of his ideas were further developed and eventually made a meaningful contribution to biology. However, his reputation was damaged to the point where a large portion of the scientific community did not take his work seriously [Shmailov, 2016].

The 1950's would pose significant problems for Rashevsky's Committee on Mathematical Biology. The cold war was in full stride and Senator Joseph McCarthy's anti-communist campaign spurred investigations in to the political leanings of academics. It was clear that Rashevsky was a not a communist, since he was a member of Ukraine's anti-communist White Guard. However, many committee members had far left ideologies and became the target of investigations. To protect such members, all individuals in the committee refused to sign loyalty oaths and Rashevsky disobeyed orders to remove specific committee members. Consequently, the University of Chicago severely cut the committee's funding and it became difficult for the group obtain grants. This challenging work environment made many

researchers transfer to other institutions and the committee was reduced to two members [Cull, 2007, Shmailov, 2016].

While certainly a setback, the committee was able to survive. Many of Rashevsky's collaborators from other institutions rallied together and published a letter in Science denouncing the Universities treatment of the committee. Additionally, Rashevsky was able to secure funding for the *Bulletin of Mathematical Biology*, ensuring the journal's survival without university support. Overtime, the program's reputation was restored and by 1960 the committee secured a 5 year NIH training grant of an amount over $500,000. This grant was a significant turning point for mathematical biology as it funded a generation of researchers and legitimized the field to other universities. As a result, universities nationwide became interested in mathematical biology and started forming their own research groups[Shmailov, 2016].

However, the committee's good fortune did not last. In the 1960s, Rashevsky was ready to retire and a new committee chairman had to be selected. Rashevsky wanted his successor to share his vision of a theoretically based mathematical biology and was adamant that the position be filled by one of his colleagues. On the other hand, the university wanted to give the committee a more experimental focus and sought out an outside hire. After several years of internal conflict, the university appointed Jack Cowan as the new chairman effective 1967. Cowan immediately changed the focus and methods of the committee. Within a year, Rashevsky's original committee was almost unrecognizable. The university even changed the committee's name to the Department of Theoretical Biology and Biophysics. The changes prompted many of the groups experienced members to leave. Additionally, several funding agencies began reevaluating the committee's new curriculum and by 1970 the committee lost most of its major grants, including the one awarded by the NIH. Eventually, Cowan transferred to the mathematics department and the committee

was disbanded[Shmailov, 2016].

The current field of biomath remembers Rashevsky with a mixed legacy. His life's work did plant the seeds that would grow into the field of biomath. Most of his work was too theoretical to be of any immediate use to scientists. However, some of his models were further developed into revolutionary ideas. For example, his early models of neural nets were developed into a form that is fundamental to artificial intelligence. Many of his students helped to establish biomath programs at other institutions and his journal, *Bulletin of Mathematical Biology*, still exists today. Regardless, the following generation of mathematical biologists felt the need to distance themselves from Rashevsky and as a result the field was renamed as "biomathematics". Rashevsky's ultimate legacy is not his research but rather his efforts to create an interdisciplinary field that uses math as a guide instead of a way to interpreted data. Rashevsky's failed research career represents a cautionary tale about the limitations of inaccessible theory. Ideally, biomath should eventually be used to inform biology. However, Rashevsky lacked the biological training needed to effectively communicate with experimentalists. This shows that the field of biomath was born with a fundamental wedge between biologists and mathematicians. Rashevsky was unable to overcome this wedge because he did not attempt to advocate his work to biologists. This shows that in order for biomath to inform biology, biomathematicians have the responsibility to reach out to experientialists and make their results relevant.

Chapter 3

Patterns of Protein Domains in Mycobacteria

This chapter will contain a section of biologically relevant background information and the write up of my summer research. The write up section is technical, however the chapter as a whole can be understood without it.

3.1 Background

Proteins are biomolecules that are often composed of multiple sub-units known as domains. A protein can be thought of a Lego model, such as a Lego car. In this metaphor, a protein domain would be a set specialized parts of the model, such as the parts that make up the wheels. Domains are especially useful because they provide a basis to compare two different proteins. For example, we can say a Lego tank is similar to a Lego car because they both have wheel domains. Similarly, a Lego tank is fairly different then a Lego boat because a boat has no wheels.

Mycobacteria are a group of bacteria that are omnipresent in the environment, but are especially common in soil and water. These bacteria are clinically relevant

and are the focus of many research groups. Mycobacteria species have been separate into two categories. One group contains all members species of the *Mycobacterium tuberculosis* complex. While the nontuberculous mycobacteria (NTM) group is composed of all other mycobacteria species. *Mycobacterium tuberculosis* has historically been a major pathogen. Additionally, recent years have seen an increase in NTM infections [Prevots and Marras, 2015]. NTMs are difficult to diagnose as they often resemble tuberculosis infections [Raju et al., 2016]. NTMs are often treated with an extended antibiotic regimen [Henry et al., 2004] which can cause harmful side effects.

3.2 Write up

3.3 Introduction

Mycobacteria pose a significant health risk worldwide, with *M. Tuberculosis* killing over a million people annually [(WHO), 2018]. Additionally, infections by nontuberculous mycobacteria (NTM) are becoming more prevalent in many regions [Prevots and Marras, 2015]. NTMs are omnipresent in the environment [Falkinham, 2009] and often act as opportunistic pathogens[Cook, 2010]. This demonstrates that mycobacteria are able to colonize a variety of environments, including the human body. It would therefore be useful to identify proteins that allow mycobacteria to survive within the human body. However, the function of many mycobacteria proteins are unknown [Kumar et al., 2017]. Therefore, identifying clinically relevant protein domains is a more feasible approach. In this paper, we study trends of domain occurrence and enrichment across 118 proteomes from a diverse set of mycobacteria species. It is our hope that this study can identify domains that would help future studies into parthenogenesis.

3.4 Methods

3.4.1 Data compilation

This work utilized the Pfam domain composition of 118 proteomes of various species of mycobacteria. A Pfam domain is a distinct functional sub-unit of a protein which is identified using a Hidden Markov Model[El-Gebali et al., 2018]. Our dataset was composed of all 117 mycobacteria proteomes listed on the main 2019 Pfam site, as well as the proteome for *M. abscessus sp. abscessus* which was retrieved from the Pfam FTP site. The domain composition of a proteome contains information on the number of sequences that correspond to a domain and the number of times each domain occurs. For this study, we only consider the number of times each domain occurs as it best reflects enrichment.

We organized the data on domain occurrences into a structure that will be referred to as the Occurrence Domain Matrix or the ODM. In the ODM, columns represent different mycobacteria proteomes and rows correspond to various domains. The matrix entries represent the number of times a particular domain occurs in the corresponding proteome. The ODM was constructed using an iterative process. The process began by adding a new column to the existing matrix to represent an additional mycobacteria proteome. Then the process compared the domains from the new proteome with the domains already present in the domain matrix. If a domain was already present, the corresponding row was updated in the new proteome column. If a domain was not already in the matrix, then a new domain row was added and the new proteome column was updated.

3.4.2 Enrichment detection

We consider a proteome to be enriched with a specific domain if it has a domain occurrence that is significantly above average. We identified enrichment by finding upper outliers in the domain occurrence from the set of all proteome that contain the domain. Outliers were determined by finding values greater than $\max\{3 \cdot S_n, 15\}$ above the median. The expression S_n, described in [Rousseeuw and Croux, 1993], is an ancillary statistic given by the formula

$$S_n = c * med_i\{med_j|x_i - x_j|\}.$$

We used S_n because it is a robust statistic that does not assume a symmetric distribution. All calculations for S_n were done using a pre-existing matlab program [Jones, 2019]. Outliers were required to be at least 15 above the median to ensure that enrichment will always imply a large disparity from the median. We consider enrichment to be significantly above average expression among all proteomes that contain a domain. Thus, we only look for outliers among the non-zero values of domain occurrence.

3.4.3 Formation of Groups

To assist with finding notable domains, we separated the proteomes into groups with similar domain compositions. We utilized a dendrogram based on Jaccard similarity indices [Beagle, 2019] and observed 44 proteomes formed a distinct cluster. The majority of proteomes in this cluster were from rapid growing species, with the exception of *M. triviale*. Therefore, we will refer to this cluster as the rapid growing group.

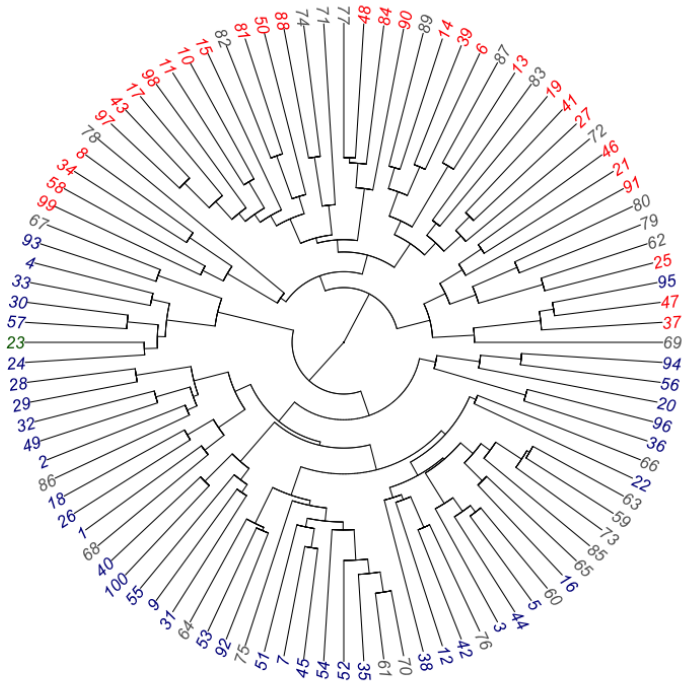

Figure 3.1: This dendrogram is constructed using Jaccard similarity indices between proteomes. Proteomes colored red correspond to rapid growing species, blue correspond to slow growing species, green is intermediate growing species, and grey is unknown.

We observed a cluster of 57 proteomes that primarily belonging to slow growing species. We will refer to this cluster as the slow growing group. We further divided the slow growing group into three subgroups based on less distinct clustering behavior. These will be referred to as slow subgroups. The slow subgroup 1 is comprised of 36 proteomes, slow subgroup 2 is made up of 13 proteomes and slow subgroup 3 is comprised of 8 proteomes. We proceed by identifying differences in domain behaviors

between these groups. There were a total of 17 proteomes that are not included in any groups. These include proteome from rapid and slow growing species. The slow group is primarily comprised of proteomes from potentially clinical species. However, slow subgroup 2 contains pathogenic species such as *M. tuberculosis*, *M. kansasii*, and *M. ulcerans*.

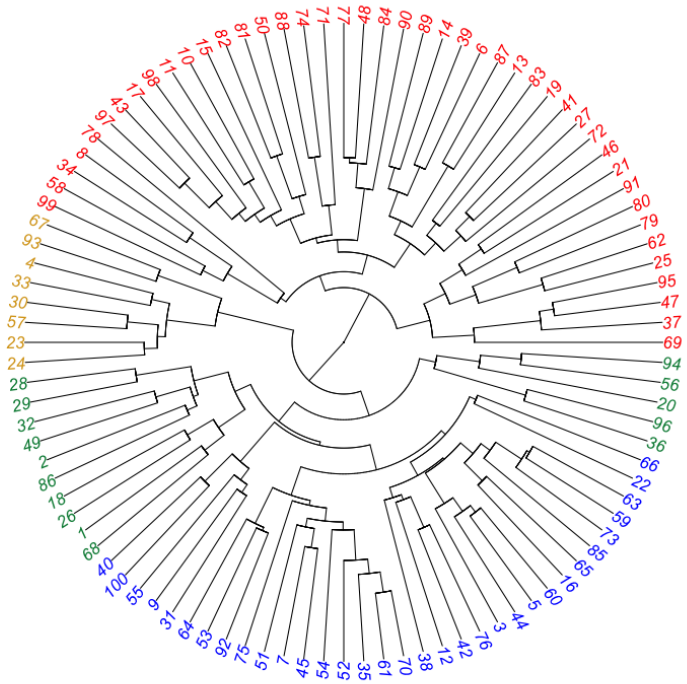

Figure 3.2: This dendrogram is constructed using Jaccard similarity indices between proteomes. Proteomes in the rapid group are colored red. Proteomes in slow subgroup 1 are colored blue. Proteomes in slow subgroup 2 are colored green. Proteomes in slow subgroup 3 are colored yellow.

3.4.4 Filtering Behaviors

We identified three commonly occurring domain behaviors between groups: exclusive domains, missing domains, and widely enriched domains. An exclusive domain is one that generally in one group and uncommon in every other group. We defined an exclusive domain as being present in 90% of proteomes in one group and at most 10% of proteomes in the other group. The subgroup clusters are less distinct from each other and thus we defined exclusive domains to be in 85% of proteomes in one group and at most 15% of proteomes in the other groups. A missing domain is one that is uncommon in one group but is generally present in every other group. Missing was defined as being in 90% of proteomes in other groups and in only 10% of proteomes in the observed group. A widely enriched domain is a one that is enriched in a 25% of proteomes in a group.

3.5 Results

Domain	RG%	SG%	SS1%	SS2%	SS3%	function
BCA_ABC_TP_C	100	0	0	0	0	transporter
Na_H_antiport_1	100	8.9	2.9	0	50	membrane
BPD_transp_2	100	5.4	5.7	7.7	0	transporter
SBP_bac_6	100	3.6	0	0	25	solute-binding
Xan_ur_permease	100	5.4	0	1.5	12.5	permeases
GntP_permease	97.7	0	0	0	0	permeases
DUF456	97.7	8.9	11.4	7.7	0	unknown
TGT	97.7	0	0	0	0	transferase
Vut_1	97.7	0	0	0	0	transporter
PepSY_TM	95.4	0	0	0	0	peptidases
PGPGW	95.4	5.4	5.7	7.7	0	transmembrane
DUF2207	93.2	8.9	5.7	0	37.5	unknown
OHCU_decarbox	93.2	3.6	0	1.5	0	enzyme
LysE	93.2	7.1	2.9	23.1	0	translocator
DUF218	93.2	3.6	5.7	0	0	unknown
PE_PPE_C	2.3	100	100	100	100	immunostimulation/virulance
Sbt_1	2.3	94.6	97.1	100	75	transporter

Table 3.1: This table shows exclusive and missing domains for each group. The middle columns represent the percentage of proteomes that contain a domain in a group or subgroup. The rightmost column describes the function of the domain's family as described by Pfam. The top section is the set of domains that are exclusive to the rapid/slow groups.

Domain	RG%	SG%	SS1%	SS2%	SS3%	function
Cyt-b5	22.7	28.6	8.6	92.3	12.5	tRNA-splicing ligase
Stealth_CR1	0	25.0	5.7	92.3	0	stealth protein
Stealth_CR2	0	26.8	8.6	92.3	0	stealth protein
TENA_THI-4	95.4	25.0	14.3	7.7	100	enhancer enzymes
ELFV_dehydrog_N	59.1	17.9	2.9	7.7	100	dehydrogenase
Ectoine_synth	61.4	17.9	5.7	0	100	ectoine synthase
DUF4126	93.2	25.0	14.3	7.7	100	unknown
DUF309	72.7	23.2	14.3	0	100	unknown
ELFV_dehydrog	59.1	17.9	2.9	7.7	100	dehydrogenase
RF3_C	93.2	12.5	0	0	87.5	release factor
GGACT	13.6	17.9	8.6	0	87.5	cyclotransferase
ScdA_N	0	12.5	0	0	87.5	repair of iron-sulphur clusters
CopC	100	71.4	91.4	0	100	blue copper protein
DUF1775	45.4	73.2	94.3	0	100	unknown
SpoIIE	100	85.7	100	100	0	sporulation protein
DUF2752	100	85.7	100	100	0	unknown
Peripla_BP_3	97.7	82.1	94.3	100	0	transcriptional regulator
LacI	97.7	80.3	94.3	92.3	0	transcriptional regulator
Peroxidase	100	85.7	100	100	0	catalyst
Voltage_CLC	95.4	80.4	91.4	100	0	chloride channel
Sulphotransf	75.0	83.9	97.1	100	0	synthesis of sulpholipid-1
DUF2277	95.4	80.4	91.4	100	0	unknown
Pro_dh	95.4	83.9	97.1	100	0	dehydrogenase
NicO	31.8	82.1	97.1	92.3	0	nickel-transport
DUF2332	22.7	78.6	91.4	92.3	0	unknown

Table 3.2: This table shows exclusive and missing domains for each subgroup. The middle columns represent the percentage of proteomes that contain a domain in a group or subgroup. The rightmost column describes the function of the domain's family as described by Pfam. The top section is the set of domains that are exclusive to the slow subgroups. The bottom section is the set of domains that are missing from among the slow subgroups.

Domain	RG%	SG%	SS1%	SS2%	SS3%	function
Big_9	38.6	0	0	0	0	diverse
BPD_transp_1	36.4	0	0	0	0	transporter
ABC_tran	31.8	0	0	0	0	transporter
PE	0	42.9	31.4	100	0	immunostimulation/virulence
Pentapeptide_2	0	19.6	0	84.6	0	unknown
PPE	0	10.7	0	46.1	0	immunostimulation/virulence
Pkinase	0	8.9	2.9	30.8	0	protein kinase
Ketoacyl-synt	0	10.7	0	46.1	0	enzyme
Acyl_transf_1	0	10.7	0	46.1	0	acyl transferase
Ketoacyl-synt_C	0	7.1	0	30.8	0	enzyme
KR	0	10.7	0	46.1	0	bacterial polyketide synthases
KAsynt_C_assoc	0	7.1	0	30.8	0	ketoacyl-synthetase
PS-DH	0	7.1	0	30.8	0	dehydratase
PIN	0	16.1	8.6	38.5	12.5	nuclease
PE-PPE	0	3.6	0	0	25	immunostimulation/virulence

Table 3.3: This table shows enriched domains for each group and subgroup. The middle columns represent the percentage of proteomes in each group or subgroup that are enriched by the corresponding domain. The rightmost column describes the function of the domain's family as described by Pfam.

3.5.1 Observations in Individual Proteomes

We looked for enrichment and unique domains among three clinically relevant proteomes of mycobacteria: *M. abscessus*, *M. avium*, and *M. tuberculosis strain ATCC 25618*. We consider a domain to be unique if it is only found in a single species of mycobacteria. Enrichment occurred in all three proteomes with *M. abscessus* being enriched with 6 domains, *M. avium* being enriched with 1 domain, and *M. tuberculosis* being enriched with 3 domains.

The average number of unique domains across proteomes is 5.2203 domains. Surprisingly, *M. abscessus* has the extreme value of 276 unique domains, while *M. avium* doesn't have any unique domains and *M. tuberculosis* has 6. Interestingly, 5 of the 6 of the unique domains in *M. tuberculosis*, i.e. Csm1_B, CRISPR_Cas6, Csm4_C, Cas_Csm6, and Csm2_III-A, are related to CRISPR systems.

Domain	*M. abscessus*	*M. avium*	*M. tuberculosis*	function
ABC_tran	126	51	46	transport
MMPL	65	38	25	integral membrane
BPD_transp_1	54	25	21	transport
Acetyltransf_1	30	10	9	transferase
Mycobact_memb	26	14	7	membrane
G5	18	1	1	extracellular
AMP-binding	69	116	43	binding
Pentapeptide_2	0	0	237	unknown
PIN	0	1	45	cleave single stranded RNA
PE	3	10	89	immunostimulation/virulence

Table 3.4: Subsection of the Occurrence Domain Matrix showing examples of enrichment across three proteomes of mycobacteria. *M. abscessus* is enriched with the domains ABC_tran, MMPL, BPD_transp_1, Acetyltransf_1, Mycobact_memb, and G5. *M. avium* is enriched with the domain AMP-binding. *M. tuberculosis strain ATCC 25618* is enriched with the domains Pentapeptide_2, PIN, and PE.

3.6 Discussion

3.6.1 Rapid Growing Group

Our results suggest that a diverse set of transport proteins may be needed to facilitate rapid growth. Many of the rapid group's exclusive and enriched domains are related to molecular transportation. Additionally, rapid growing species had a relatively large amount of exclusive domains but few enriched domains

3.6.2 Slow Growing Group

The slow grouping group had few exclusive or enriched domains. However, the only domain to be shared by the entire group is PE_PPE_C, which may have a function related to virulence and immunostimulation. Having commonly shared immunostimulation protein domains may explain the opportunisticly pathogenic behavior of slow growing mycobacteria.

3.6.3 Slow Subgroup 1

Slow subgroup 1 has very little interesting domain behavior. This potentially could be the result of the subgroup including a relatively large amount of proteomes. Additionally, exclusive and missing behaviors may be difficult to detect because this subgroup is closely related to slow subgroup 2.

3.6.4 Slow Subgroup 2

Slow subgroup 2 includes many clinical species so it's domain behavior is particularly relevant. The group exclusively shared stealth proteins which help to evade immune systems. Additionally, slow group 2 is widely enriched with PPE and PE, both of which may have a function related to virulence. It is worth noting that this subgroup experienced the most widely enriched domains.

3.6.5 Slow Subgroup 3

Slow subgroup 3 had a lot of exclusive and missing domain while featuring limited widespread enrichment. The subgroup is comprised of proteomes from both environmental and opportunistically pathogenic species. Since the majority of the slow growing proteomes come from opportunistically pathogenic species, it is unlikely that any of the subgroup 3 exclusive or missing domains have clinical relevance.

3.7 Acknowledgements

This research was conducted as part of Colorado Biomedical Informatics Summer Training Fellowship at the University of Colorado Anschutz Medical Campus in summer 2019, funded by T15 Training Grant no.-2T15LM009451. Thank you to Michael

Strong for his mentorship as well as Sean Beagle for offering insights into this work.

3.8 Overview

This work can be considered an exploration into bioinformatics. My general approach was to collect all available data from a database and look for patterns. More specifically, I looked for domains that were only found in slow growing and fast growing species. One main finding was that slow growing and fast growing mycobacteria generally have different domain compositions. Additionally, I was able to produce a list of domains that were exclusive to either slow and fast growing species. This research showed me the value of understandings the biological data I was working with. At the time of this work, I did not understand mycobacteria protein domains enough to interpret my own result. For example, I was not sure which of the domains on my exclusive and enriched list were interesting to talk about. Luckily, my research mentor was able to look at my list and suggest a few talking points for me. However, it was strange that I was not able to understand the result on my own project.

Chapter 4

Modern Biomathematics

Mathematical modeling essentially describes a biological phenomenon by using a set of assumptions to replicate an observed behavior. This approach has led to success in several research areas where direct experimentation is not possible due to logistical restraints. Models are usually created as a collaboration between mathematicians and biologists. However, collaboration can cause miscommunications between experts which results in low quality models.

Evolution is a mechanism that shapes the development of all life. However, it is a multigenerational process that usually takes too long to study directly. As a result, mathematical models based on game theory have been particularly usual in studying the evolutionary mechanisms that maintain the existence of specific traits. Game theory is the mathematics of decision making and it assumes individuals are rational and self-interested. In a biological context, it has been used to explain why certain behaviors are evolutionary beneficial [Smith and Price, 1973]. This is done by considering how a set of theoretical animals interact with the general goal of finding a set of assumptions that encourage the virtual animals to behave the same way they do in nature. Assumptions are incorporated in the form of various strategies animals use to compete with each other. This approach has offered insight into a variety of

biological phenomena, including the evolution of cooperation [Nowak, 2012].

However, a mathematical model is only as good as the assumptions used to build it, and a lack of biological understanding can easily lead a model astray. One of the early models for HIV transmission was overly complicated and built on inaccurate assumptions about the disease [May, 2004]. Collaborations between mathematicians and biologists are not enough to prevent such incidents because the assumptions may be too mathematical for biologist to notice. In the case of the HIV model, researchers were repurposing an existing mathematical model for measles. However, they did not understand that the two viruses spread in fundamentally different ways. Therefore, they assumed that an individual who was having sex with a new partner 10 times was just as likely to catch HIV as an individual who was having sex with 10 new partners one time. Additionally, it would have been difficult for a biologist to catch such a mistake because the model was build using advanced mathematical tools, such as partial differential equations, that biologists are not trained to understand. Having researchers adequately trained in both math and biology would prevent such incidents.

An increasingly prominent approach in mathematical biology is using data to drive investigations. This version of mathematical biology tries to limit assumptions about a system and let data drive discovery. This is the data focused approach I used in Chapter 2 and has been utilized in many areas of biology including bioinformatics, systems biology, and quantitative ecology. The general goal of a data driven approach is to make sense of a data set by looking for correlations or interesting patterns. It is common for researchers in this area to utilize data that experimentalists have previously collected.

This method is particularly effective at interpreting large data sets. In Chapter 2, I identified interesting protein domains by filtering for domains that were in some

species but not others. My data was composed of 118 species and included over 4000 domains. It would be difficult for a biologist to investigate how all of these domains are used in each species. However, my work was able to filter these domains into a list that is small enough that a biologist would be able interpret noteworthy results.

As a consequence of using existing data sets, researchers may not detect or understand potential limitations of their work. In my domain work, I trusted that the PFAM database had quality data. However, the phenotypic information of several species was limited. This means that I had data about which domains that occur in each species of mycobacteria but I do not know any of the traits, such as where a strain found and the species growing rate, and I was only able to make limited conclusions about domains. Additionally, I am not sure how PFAM selects species to include in its database, consequently I could have accidentally used a biased data set.

Perhaps, the biggest difference between a mathematical model and a data driven approach is the way biologists engage with the results. Statistics report results with common metrics such as p-values, confidence intervals, and R values. This is convenient because scientists with limited training in statistics are still able to understand the result of a wide variety of statistical analysis. For example, I was able to explain the work done in Chapter 2 to an expert on mycobacteria and he was able to interpret my results for me. However, mathematical models can be based on a wide variety of mathematics and thus a standard metric is not possible. For example, ecological models based on the advanced mathematical techniques of partial differential equations have been effectively used to model the spread of an invasive species [Holmes et al., 1994]. The classical model was used to describe the spread of muskrats populations through Europe [Skellam, 1951]. The model used in [Skellam, 1951], is of the form

$$p = \int_R^\infty \exp[-r^2/na^2]2rdr/na^2 = \exp\{-R^2/na^2\}.$$

This expression is clearly complicated and any useful interpretation requires a background in partial differential equations. This demonstrates that biological systems are often complicated and quality models incorporate advanced mathematics. The results of such models can be particularly difficult to interpret because the model's results demands a mathematical background while the contexts needs a biological understanding. Therefore, the researchers interpreting such models need a strong background in both mathematics and biology.

Chapter 5

Benefits of integration

Modern biomath is as a collaboration between biologists and mathematicians. This works well for specific problems. However, the stark divorce of expertise in mathematical tools and biological is limiting the potential of biomath. Therefore, it is necessary to train biomathematicians in advanced mathematics and biology. It would be unreasonable to train all biologists in advanced mathematics. Consequently, biomath will always be a type of specialization. Treating biomath as a subfield within biology would increase the accessibility and overall impact of the field.

In light of this, it may be worthwhile to reconsider Rashevsky's vision for a new type of biologists. In fact, many contemporary scholars argue that mathematical biology should be analogous to theoretical physics [van Hemmen, 2007]. The main drivers of mathematical biology needs to be people who understands both the biology involved in a system and the mathematical tools available.

Having researchers with an adequate understanding of both biology and mathematics would allow modelers to interpret their own results. A lot of mathematical models are difficult to interpret without the proper training. It is not always possible to find a biologist with the training to read a purposed model and interpret the results biologically. Therefore, it is the responsibility of the modeler to interpret their own

results and understand what they reveal about a biological system.

Rashevsky's career highlights the need for a modeler to effetely communicate with the biological community. A mathematical biologists who trained in both math and biology would be able to facilitate communication between the two fields. This is valuable because advocating for a mathematical model should include showing why the work is sound mathematically and informative biologically. Therefore, modelers need to converse with and justify themselves to experts in both areas.

Many areas of biology, such as molecular biology, are developing very quickly and modelers need the ability to read new publications to ensure they are building models with current information. This may seem trivial however, most biological journal publications are written with the assumption that the reader have a basic understanding of the biology. Therefore, mathematicians without any biological training would have a difficult time fully understanding the information being presented.

Mathematical models are useful because they can be repurposed to describe a variety of biological systems. For example, many of the current mathematical models that are used to describe invasive species are variations of a single original model [Skellam, 1951]. Such expansions would have been possible if biologists were not able to recognize how the model could be adapted to their own work. Therefore, model results should be accessible to general biologists, which is only possible if modelers are able to effectively communicate their methods and results.

This work may seem only relevant to biologists and some mathematicians. However, we are living beings and therefore biology affects us all. If we can use mathematical models to better study biology then we gain a better understanding of ourselves and the world around us. After all, models can potentially be used to study topics relating to human health, global warming, and many more of the things that affect us all.

Bibliography

[Abraham, 2004] Abraham, T. H. (2004). Nicolas rashevsky's mathematical biophysics. *Journal of the History of Biology*, 37(2):333–385.

[Beagle, 2019] Beagle, S. (2019). Dendrogram. `https://github.com/SeanBeagle/Dendrogram`.

[Cook, 2010] Cook, J. L. (2010). Nontuberculous mycobacteria: opportunistic environmental pathogens for predisposed hosts. *British medical bulletin*, 96(1):45–59.

[Cull, 2007] Cull, P. (2007). The mathematical biophysics of nicolas rashevsky. *Biosystems*, 88(3):178–184.

[El-Gebali et al., 2018] El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., et al. (2018). The pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432.

[Falkinham, 2009] Falkinham, III, J. (2009). Surrounded by mycobacteria: nontuberculous mycobacteria in the human environment. *Journal of applied microbiology*, 107(2):356–367.

[Henry et al., 2004] Henry, M., Inamdar, L., O'riordain, D., Schweiger, M., and Watson, J. (2004). Nontuberculous mycobacteria in non-hiv patients: epidemiology, treatment and response. *European Respiratory Journal*, 23(5):741–746.

[Holmes et al., 1994] Holmes, E. E., Lewis, M. A., Banks, J., and Veit, R. (1994). Partial differential equations in ecology: spatial interactions and population dynamics. *Ecology*, 75(1):17–29.

[Jones, 2019] Jones, P. (2019). Rousseeuwcrouxsn. `https://github.com/petejonze/psychosim/blob/master/RousseeuwCrouxSn.m`.

[Kumar et al., 2017] Kumar, A., Sharma, A., Kaur, G., Makkar, P., and Kaur, J. (2017). Functional characterization of hypothetical proteins of mycobacterium tuberculosis with possible esterase/lipase signature: a cumulative in silico and in vitro approach. *Journal of Biomolecular Structure and Dynamics*, 35(6):1226–1243.

[Mackey and Maini, 2015] Mackey, M. C. and Maini, P. K. (2015). What has mathematics done for biology? *Bulletin of mathematical biology*, 77(5):735–738.

[May, 2004] May, R. M. (2004). Uses and abuses of mathematics in biology. *Science*, 303(5659):790–793.

[Nowak, 2012] Nowak, M. A. (2012). Evolving cooperation. *Journal of theoretical biology*, 299(0):1–8.

[Prevots and Marras, 2015] Prevots, D. R. and Marras, T. K. (2015). Epidemiology of human pulmonary infection with nontuberculous mycobacteria: a review. *Clinics in chest medicine*, 36(1):13–34.

[Raju et al., 2016] Raju, R. M., Raju, S. M., Zhao, Y., and Rubin, E. J. (2016). Leveraging advances in tuberculosis diagnosis and treatment to address nontuberculous mycobacterial disease. *Emerging infectious diseases*, 22(3):365.

[Rosen, 1991] Rosen, R. (1991). *Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life*. Columbia University Press.

[Rousseeuw and Croux, 1993] Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283.

[Shmailov, 2016] Shmailov, M. M. (2016). *Intellectual Pursuits of Nicolas Rashevsky: The Queer Duck of Biology*, volume 55. Birkhäuser.

[Skellam, 1951] Skellam, J. G. (1951). Random dispersal in theoretical populations. *Biometrika*, 38(1/2):196–218.

[Smith and Price, 1973] Smith, J. M. and Price, G. R. (1973). The logic of animal conflict. *Nature*, 246(5427):15.

[van Hemmen, 2007] van Hemmen, J. L. (2007). Biology and mathematics: A fruitful merger of two cultures. *Biological cybernetics*, 97(1):1–3.

[(WHO), 2018] (WHO), W. H. O. (2018). Tuberculosis. `https://www.who.int/gho/tb/epidemic/cases_deaths/en/`.