# The Effects Of CRISPRs On Horizontal Gene Transfer: A Network Theoretic Approach

MolBiol 4C12 Research Proposal

Siddharth Reed*[1] and G. Brian Golding[1]

[1]Department of Biology, McMaster University
Hamilton,Canada

January 22, 2019

**Abstract**

Horiznotal Gene Transfer (HGT) is a mechanism by which organisms (mainly prokaryotes) can share genetic material outside of inheritance. HGT has proven to have significant effects on bacterial genome evolution, allowing for increased genetic diversity and niche adaptation. CRISPR associated (CRISPR-Cas) is an adaptive immune system in prokaryotes that has garnered much research attentions due to it's applications as a gene editing tool. While much of the focus on CRISPR-Cas systems has been related to this application, CRISPR-Cas has been shown to have a highly complex interaction with HGT. Effort has been focused on the how CRISPR-Cas systems affect the mechanisms of HGT, little is currently known about the effects of CRISPR-Cas on HGT rate, both for individuals and on a population level. Here is proposed a network-theoretic approach to further the understanding of the effects of the presence of CRISPR-Cas systems on HGT within a population. Network theory has already been applied to better model evolution and relatedness among bacteria, accounting for HGT which traditional phylogenetic methods ignore. This network-theoretic approach allows for study of CRISPR-Cas effects on individual bacteria as well as population level effects on HGT. Understanding the effects of CRISPR-Cas on HGT may help develop strategies to curb spreading antibiotic resistance, understanding bacterial evolution and extend the functionality of CRISPR-Cas ene editing systems.

*To whom correspondence should be addressed; reeds4@mcmaster.ca

# 1 CRISPR-Cas Systems

## 1.1 What Are They?

CRISPR-Cas systems are sets of nucleotide motifs (spacers) interspaced with nucleotide repeats (Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)s) and CRISPR-associated (Cas) proteins (usually adjacent to the CRISPR motifs) that serve an adaptive immune function in many bacteria and archaea[27]. Each nucleotide motif is indicative of some DNA sequence that was taken up previously by the host and serves as a marker for the Cas proteins to degrade any foreign DNA matching this motif[27]. For example, if a bacterium is infected with a phage and survives, a motif representative of that phage can be integrated into the CRISPR repeats so that when the phage attempts infection again, it will be degraded by a Cas protein before genomic integration can occur. CRISPR is an adaptive immune systems, as it adapts resistance after an initial infection, which is generally required for the spacer sequence to be integrated.

Despite seeming to function primarily as an immune system, non-viral spacers representative of bacterial Mobile Genetic Element (MGE)s have been found to compose a majority of (detectable) CRISPR spacers[30]. In fact, many spacers were found to have no detectable match, being termed CRISPR "dark matter", indicating that knowledge about the acquisition of spacers and their effects on bacterial gene dynamics leaves much to be desired[30].

## 1.2 Diversity, Ubiquity And Detection

As of 2017, over 45% of bacterial genomes analyzed ($n = 6782$) have been found to have CRISPR motifs[11]. CRISPR motifs show significant diversity between organisms, as expected since they represent a chronological history of viral infection or MGE "infection"[27]. But even Cas proteins themselves also show diversity and evolution, segregating into entirely different CRISPR-Cas systems[18]. There still exist many bacterial strains, and even entire genera with no known CRISPR-Cas systems, although the emphasis should be placed on the word *known*[37, 13].

Between $11\% - 28\%$ of sequenced genomes have either only CRISPR repeats *or* Cas loci, but not both[37]. There also exist repeat motifs that may superficially resemble CRISPRs, but have low spacer diversity and no Cas genes[37]. False detection of CRISPR systems is significantly increased by only considering repeat-spacer structural patterns, and considerations of more information, such as spacer dissimilarity and genomic context are necessary for reducing false positives[37]. Especially as sequencing efforts continue, better mechanistic understandings of CRISPR systems develop and CRISPR systems themselves propagate and transfer between bacteria they will continue to become more relevant and diverse[27].

## 1.3 Applications In Biotechnology

While CRISPRs are interesting systems to study from a microbiological perspective, much of the current research interest (and funding) is motivated by applications of CRISPR to gene editing. CRISPR-Cas9 has been developed into a simple, efficient tool for gene editing in

both prokaryotes and eukaryotes[27]. Using the Cas-9 protein which induces a double-strand break to a region homologous to a specified guide RNA. The break will then re-anneal (likely) incorrectly, removing gene function[27]. A gene can also be inserted at the break point via homology directed repair by including DNA sequence with flanking arms homologous to the break region[27].

# 2 Do CRISPR Systems Affect Horizontal Gene Transfer?

Yes.

## 2.1 Interference Mechanisms

Since CRISPRs have been shown to be capable of interfering with conjugation (conjugative plasmid specific spacers) and transduction (phage immunity), it has been hypothesized that lower rates of HGT will be observed in strains with CRISPR-Cas systems[19]. CRISPR-Cas systems have also been found to interfere with transformation-mediated HGT, by degrading foreign DNA taken up by a cell[38].

## 2.2 Complexities And Costs Of CRISPR-Cas Systems

As noted above CRISPR-Cas systems have been shown to interfere with plasmid conjugation in *Staphylococcus aureus* (*S. aureus*) by integrating a spacer targeting the *nickase* sequence, necessary for conjugation in *S. aureus*[19]. Since antibiotic resistance genes are often transferred on plasmids, this can incur a significant cost, especially in environments with large amounts of antibiotics (ex: hospitals, trees etc.)[8]. CRISPR-Cas systems incur a metabolic cost, as Cas proteins, guide RNAs, spacer acquisition proteins must all be expressed to maintain immunity[27]. Despite primarily being an immune system, the way CRISPR-Cas functions (degrading foreign DNA matching spacers motifs, resisting phage infection) can have off-target effects on HGT[5]. While resisting lytic phage infection clearly provides some fitness benefit, CRISPR-Cas has also been shown to resist prophage incorporation[5]. Prophages can serve as vectors for HGT, but they can also provide super-infection immunity, and even reduce competitor bacterial populations through infection[5, 34]. It has also been shown that spacer sequences representative of a bacterium's own chromosomal DNA can be incorporated in to CRISPR array, leading to an auto-immune response where Cas proteins target native host DNA[31]. As CRISPR-Cas systems persist, anti-CRISPR mechanisms have evolved in certain phages, making them immune to CRISPR-Cas, denoted anti-CRISPRs[5]. This has a two-fold effect, as it can increase the susceptibility of the host to infection, reducing the fitness benefit of CRISPR-Cas, but it can also allow for more transduction-mediated HGT[5].

### 2.3 Potential Strategies For Reducing CRISPR-HGT Trade-off Costs

Due to the myriad of fitness costs associated with consistently expressing CRISPR-Cas systems, bacteria have appear to develop strategies to mitigate these costs. While CRISPR-Cas system can confer a fitness advantage by providing immunity to phage infection, the fitness cost associated is complex, especially as CRISPR-Cas systems themselves can be transferred horizontally, either on a plasmid or even through transduction[9]. It has been posited that CRISPR-Cas systems need only be present in a few members of a population at once and transferred between members to maintain phage immunity while reduce the cost of constantly maintaining CRISPR-Cas systems[5]. It has been found that the presence of a CRISPR system does not necessarily imply activity of the system, creating another mechanisms by which the fitness cost of CRISPR-Cas systems can be reduced[5]. The presence of CRISPR-Cas systems have also been shown to actually enhance HGT via transduction at the population level by reducing total phage abundance[34]. The presence of CRISPR-Cas systems in Firmicutes have been shown to be associated with increased levels of gene insertion and deletion compared to closely related outgroups, further demonstrating the complexity of this relationship[36]. What is clear is that the effects of CRISPR-Cas systems on rates of HGT are highly complex. This is owning in no small part to the broad range of CRISPR effect, how CRISPR activity can be modulated and the transfer of CRISPR systems themselves within a population[5]. Taking a systematic approach may help elucidate the dynamics between CRISPR system presence and HGT rate.

## 3 Horizontal Gene Transfer

HGT can be defined as the exchange of genetic information across lineages[39]. The word horizontal is in contrast to what can be referred to as "vertical" gene, between parents and offspring[28]. HGT is often a source of novel adaptations, allowing organisms to respond to selective pressures much more quickly than having to evolve new functions in genes themselves[28, 20].

### 3.1 Mechanisms

**Transformation** Transformation is the uptake of free floating exogenous DNA by a bacterium and incorporates it into it's genome[39]. Many factors can influence the competency (capability of transformation) of bacteria naturally, such as DNA damage, selective pressures, cell density and multiple methods have been found to induce competency for experimental purposes (cloning)[4].

**Conjugation** Conjugation is the sharing of genetic material through cell-to-cell bridges, usually carried on either a self-transmissible or non-self-transmissible plasmid[7].

**Transduction** Transduction is the transfer of genes between bacteria through a bacteriophage[10]. When a donor cell infected by a phages is lysed, the lysed bacterial fragments

can accidentally be incorporated into the phage head[10]. When the phage infects a new bacterium the lysed donor fragments are released into the recipient cell, where they can recombine in to the genome[10]. While random (general) gene fragments can be transferred as motioned above another type of transduction exists for lysogenic phage[10]. Lysogenic phage incorporate themselves into specific regions of a bacterial genomes[10]. When they excise themselves they can accidentally incorporate bacterial DNA flanking the incorporated phage DNA and bring it with them to the next page target[10].

It should be noted that *successful* HGT requires that a gene be maintained, either by genomic integration or plasmid replication. Frequently putatively transferred genes are either lost quickly after transfer or evolve with little functional constraint, as no selective pressure is maintaining them (presumably)[14].

## 3.2 Rate Influencing Factors

The rate of HGT in bacteria is constantly in flux, in part due to the amount of DNA available for transfer[25]. If there are low levels of exogenous DNA, low population density or low phage density, reduced HGT will be observed as less DNA available for transfer[39]. But just like mutation rates, HGT rates are thought to evolve in response to environmental factors or selective pressure[35, 21]. For example, for strains of bacteria in hospitals, the potential benefit of receiving antibiotic resistance genes via HGT may far outweigh any potential danger or metabolic cost, inducing a response increasing a bacterium's uptake of foreign DNA.[8] There are clear metabolic costs for HGT, as host machinery to allow competency and conjugation are not trivial to synthesize[2]. Further, conjugation and transformation are not discriminatory processes, so DNA encoding for toxic products, having sub-optimal codon distribution or incompatible GC content may be taken up, but cannot be successfully incorporated or consistently expressed[2]. Conjugated plasmids may also be incompatible with a host due to the replication machinery required by the plasmid[24]. In fact it has been suggested that genes recently acquired via HGT are often quickly lost, having been lost for conferring no advantage or for conferring a specific advantage, that was lost with the removal of whatever selective pressure was maintaining it[14]. Ultimately HGT rates are influenced by a variety of factors, related to barriers and fitness costs/benefits associated with the genes being gained.

## 3.3 Pan-genomes

As sequencing costs has decreased, re-sequencing of strains and sequencing of many similar strains has grown drastically. From this different strains of the same species were compared yielding interesting results: many genes are not found in most of the strains sequenced[29]. This has lead to the concept of a pan-genome, the sum total of all unique genes among a set of strains[12]. A pan-genome has two parts: a core genome consisting of genes common to all strains in a species and an accessory genome, consisting of unique genes present in any of the strains[12]. In *Escherichia coli* (*E. coli*), it appears that as the total number of strains sequenced increases, the total number of unique genes increases logarithmically, meaning more unique genes are being identified with every new strain sequenced[26]. These

accessory genes are prime candidates for HGT, as there are strains known not to have them and may provide some niche-specific adaptation, such as antibiotic resistance[12]. The accessory genome can be considered a genetic toolbox that strains have access to through HGT, although this access is also limited by barriers to HGT (ex: distance, genetic incompatibility etc.). Pan-genomes can further be categorized as open, if they appear to be expanding, adding more genes from more distant Operation Taxonomic Unit (OTU)s or closed, with the total number of unique genes plateauing as more strains are sequenced[12].

### 3.4 Applications

While the vast majority of HGT has been observed to occur between prokaryotes, cases have been identified between prokaryotes and eukaryotes. One particular case allowing a beetle to colonize coffee beans after (presumably) gaining a gene from a bacterial strain colonizing it's midgut[1]. This has become an issue as this beetle is considered a pest, estimated to be the cause of over 20 million USD loses to rural farming families internationally[1].

Another much more important reason for studying HGT is that it is antibiotic resistance genes have been shown to transfer frequently[33]. The transfer of antibiotic resistance genes is so prevalent that the term resistome has be coined to refer to the set of resistance genes that an organism can acquire via HGT[33]. Understanding the dynamics of HGT and ways to limit or inhibit it specifically may prove integral to resolving the issue of the decreasing range of antibiotic effectiveness[33].

# 4 Hypothesis

The null hypothesis is that bacterial strains/genera with known CRISPR systems will show no significant differences in network statistics to those strains/genera without known CRISPR systems.

# 5 Methods

### 5.1 Data Collection

Complete genomes from NCBI RefSeq are downloaded and the CRISPRdb (along with a python script) is used to annotate genera as being mixed (containing strains with and without CRISPR-Cas systems) or Non-CRISPR (containing no strains with a CRISPR-Cas system)[11]. CRISPR annotations of Cas,Cfp proteins from NCBI and the CRISPRone tool from Zhang and Ye will also be used to assess the presence of CRISPR systems[37].

### 5.2 Gene Presence/Absence Matrix

In order to use the program markophylo to estimate insertion and deletion rates, a Presence/Absence (P/A) matrix and a phylogenetic species tree are required. First any genes

classified as MGEs (from NCBI annotations) are removed. Next genes are grouped into families by reciprocal BLAST hits and single link clustering. Genes not placed in any family this way are BLASTed against the NCBI NR database to check if they are valid genes, if they are they are consider their own family. The P/A matrix is constructed as follows, for each OTU a binary vector is created, where each entry represent a gene family and a 1 indicate that that OTU contains 1 gene in that family. This is repeated for all OTU, creating a $G \times O$ binary matrix, where $G$ is the total number of gene families and $O$ is the number OTUs.

There are many ways to construct a species tree, but for this project the tree will be constructed using genes from gene families present in all OTUs being considered, using Bayesian methods, as implemented in the program MrBayes.

## 5.3 Makophylo Rate Estimations

Given a species tree and a gene family P/A matrix for the OTUs of the species tree the R package *markophylo* can provide gene insertion and gene deletion rate estimates[6]. The presence or absence of gene families are considered 2 discrete states, for which a $(2 \times 2)$ transition rate matrix (of a Continuous-time Markov cain with finite state space (CTMC-FFS) model) can be estimated using maximum likelihood techniques. This values in this estimated transition matrix are the insertion rate (transition probability of gene absence $\rightarrow$ presence)and deletion rate (transition probability of gene presence $\rightarrow$ absence)[6].

## 5.4 Network Construction

Quartet decomposition is method by which HGT events can be identified using a set of gene trees and a species tree. Given a tree $T$ a quartet is a subtree contain 4 of the leaf nodes in $T$, meaning that for a tree with $N$ leaf nodes (or OTUs) there are $\binom{N}{4}$ unique quartets in that tree. A quartet $Q$ is considered consistent with a tree if $Q = T|Le(Q)$ where $T|Le(Q)$ is the tree obtained by suppressing all degree-two nodes in $T[X]$ and $T[X]$ is the minimal subtree of T with all nodes in $X$, which is a leaf set of $T$[3]. To calculate the weight of an edge for the network, given a species tree $S$ and a set of gene trees $G$[3]:

1. Pick a horizontal edge $H = ((u, v), (v, u))$ from $S$
2. Pick a gene tree $G_i$ in $G$
3. Decompose $G_i$ into it's set of quartets $\phi_i$
4. Remove all quartets consistent with $S$ or previously explained from $\phi_i$
5. Set $RS((u, v), \phi_i)$ to be the number of quartets in $\phi_i$ that support the edge $(u, v)$
6. Set $NS((u, v), \phi_i)$ to be $RS((u, v), \phi_i)$ divided by $\lambda$, which is the total number of quartets in $S$ that are consistent with the edge $(u, v)$.
7. The score for the edge $H$ for tree $G_i$ is $max\{NS((u, v), \phi_i), NS((v, u), \phi_i)\}$
8. The total score for the edge $H$ is the sum of scores for each tree $G_i$

9. This total score calculation is repeated for each horizontal edge $H_i$ in S, resulting in a list of edges, which is a complete description of the network.

### 5.5 Network Statistics

All networks will be comprised of nodes representing OTUs and weighted edges represent the estimated amount of HGT events between the two incident nodes. As multiple set of networks can be computed for a single set of genera (using different sets of gene trees), bootstrap support for edges and confidence intervals on edge weights can also be calculated. Given a network, with a set of nodes $V = \{V_0 \dots V_i\}$ of cardinality $N$ and a set of weighted edges (an unordered 2-tuple and weight) $T = \{((V_1, V_2), W_{1,2}) \dots ((V_i, V_j), W_{i,j})\}$ with cardinality $E$ descriptive statistics can be computed as follows[22]:

- Total edge weight: sum of all edge weights in a network
- Average edge weight: sum of all edge weights divided by $N$
- Node Closeness Centrality: $\frac{N-1}{\sum_v d(x,v)}$ where $d(x, y)$ is the length of the shortest path between node v and x.
- Node Associativity: $\frac{j(j+1)(\overline{k}-\mu_q)}{2E\sigma_q^2}$ where $j$ is the excess degree of the node and $\overline{k}$ is the average excess degree of the node's neighbors and $\mu_q$ and $\sigma_q$ are the mean and standard variation of the excess degree distribution.
- Network Density: $\frac{2(E-N+1)}{N(N-3)+2}$
- Node Clustering Coefficient: $\frac{2e}{k(k-1)}$ where $k$ is the number of neighbors and $e$ is the number of edges between all neighbors.
- Network Diameter: The shortest path between the 2 furthest nodes in a network.

# References


[1] Ricardo Acuña et al. "Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee". In: *Proceedings of the National Academy of Sciences* (2012). ISSN: 0027-8424. DOI: 10.1073/pnas.1121190109. eprint: http://www.pnas.org/content/early/2012/02/17/1121190109.full.pdf. URL: http://www.pnas.org/content/early/2012/02/17/1121190109.

[2] David A. Baltrus. "Exploring the costs of horizontal gene transfer". In: *Trends in Ecology and Evolution* 28.8 (2013), pp. 489–495. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2013.04.002. URL: http://www.sciencedirect.com/science/article/pii/S0169534713001043.

[3] Mukul S. Bansal et al. "Systematic inference of highways of horizontal gene transfer in prokaryotes". In: *Bioinformatics* 29.5 (2013), pp. 571–579. DOI: 10.1093/bioinformatics/btt021. eprint: /oup/backfile/content_public/journal/bioinformatics/29/5/10.1093_bioinformatics_btt021/2/btt021.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btt021.

[4] Melanie Blokesch. "Natural competence for transformation". In: *Current Biology* 26.21 (2016), R1126–R1130. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2016.08.058. URL: http://www.sciencedirect.com/science/article/pii/S096098221631003X.

[5] J. Bondy-Denomy and A. R. Davidson. "To Acquire Or Resist:The Complex Biological Effects Of CRISPR-Cas systems". In: *Trends Microbio.* 22.4 (Apr. 2014), pp. 218–25. DOI: 10.1016/j.tim.2014.01.007.

[6] Utkarsh J. Dang and G. Brian Golding. "markophylo: Markov chain analysis on phylogenetic trees". In: *Bioinformatics* 32.1 (2016), pp. 130–132. DOI: 10.1093/bioinformatics/btv541. eprint: /oup/backfile/content_public/journal/bioinformatics/32/1/10.1093_bioinformatics_btv541/3/btv541.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btv541.

[7] J. Davison. "Genetic exchange between bacteria in the environment". In: *Plasmid* 42 (1999), pp. 73–91.

[8] Senka Dzidic and Vladimir Bedeković. "Horizontal gene transfer-emerging multidrug resistance in hospital bacteria". In: *Acta pharmacologica Sinica* 24.6 (June 2003), pp. 519–526. ISSN: 1671-4083. URL: http://www.chinaphar.com/1671-4083/24/519.htm.

[9] James S. Godde and Amanda Bickerton. "The Repetitive DNA Elements Called CRISPRs and Their Associated Genes: Evidence of Horizontal Transfer Among Prokaryotes". In: *Journal of Molecular Evolution* 62.6 (June 2006), pp. 718–729. ISSN: 1432-1432. DOI: 10.1007/s00239-005-0223-z. URL: https://doi.org/10.1007/s00239-005-0223-z.

[10] A. J. F. Griffiths et al. *An Introduction to Genetic Analysis* $7^{th}$ *Edition*. W.H. Freeman, 2000.

[11] GRissa, I. and Drevet, C. and Couvin, D. *CRISPRdb*. http://crispr.i2bc.paris-saclay.fr/. Online; accessed 22 October 2018. 2017.

[12] L. C. Guimaraes et al. “Inside the Pan-genome - Methods and Software Overview”. In: *Curr. Genomics* 16.4 (Aug. 2015), pp. 245–252.

[13] D. H. Haft et al. “A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes”. In: *PLoS Comput. Biol.* 1.6 (Nov. 2005), e60. DOI: https://doi.org/10.1371/journal.pcbi.0010060.

[14] W. Hao and G. B. Golding. “The fate of laterally transferred genes: life in the fast lane to adaptation or death”. In: *Genome Res.* 16.5 (May 2006), pp. 636–643.

[15] G. Hickey et al. “SPR distance computation for unrooted trees”. In: *Evol. Bioinform. Online* 4 (Feb. 2008), pp. 17–27.

[16] V. Kunin et al. “The net of life: reconstructing the microbial phylogenetic network”. In: *Genome Res.* 15.7 (July 2005), pp. 954–959.

[17] C.G. Kurland. “Codon bias and gene expression”. In: *FEBS Letters* 285.2 (), pp. 165–169. DOI: 10.1016/0014-5793(91)80797-7. eprint: https://febs.onlinelibrary.wiley.com/doi/pdf/10.1016/0014-5793%2891%2980797-7. URL: https://febs.onlinelibrary.wiley.com/doi/abs/10.1016/0014-5793%2891%2980797-7.

[18] K. S. Makarova et al. “Evolution and classification of the CRISPR-Cas systems”. In: *Nat. Rev. Microbiol.* 9.6 (June 2011), pp. 467–477.

[19] Luciano A. Marraffini and Erik J. Sontheimer. “CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA”. In: *Science* 322.5909 (2008), pp. 1843–1845. ISSN: 0036-8075. DOI: 10.1126/science.1165771. eprint: http://science.sciencemag.org/content/322/5909/1843.full.pdf. URL: http://science.sciencemag.org/content/322/5909/1843.

[20] P. R. Marri, W. Hao, and G. B. Golding. “The role of laterally transferred genes in adaptive evolution”. In: *BMC Evol. Biol.* 7 Suppl 1 (Feb. 2007), S8.

[21] Vadim Mozhayskiy and Ilias Tagkopoulos. “Horizontal gene transfer dynamics and distribution of fitness effects during microbial in silico evolution”. In: *BMC Bioinformatics* 13.10 (June 2012), S13. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-S10-S13. URL: https://doi.org/10.1186/1471-2105-13-S10-S13.

[22] M. Newman. “The Structure and Function of Complex Networks”. In: *SIAM Review* 45.2 (2003), pp. 167–256. DOI: 10.1137/S003614450342480. eprint: https://doi.org/10.1137/S003614450342480. URL: https://doi.org/10.1137/S003614450342480.

[23] M. E. J. Newman. "Assortative Mixing in Networks". In: *Phys. Rev. Lett.* 89 (20 Oct. 2002), p. 208701. DOI: 10.1103/PhysRevLett.89.208701. URL: https://link.aps.org/doi/10.1103/PhysRevLett.89.208701.

[24] R.P. Novick. "Plasmid Incompatibility". In: *Microbiol Rev* 51.4 (Dec. 1987), pp. 381–95.

[25] Ovidiu Popa and Tal Dagan. "Trends and barriers to lateral gene transfer in prokaryotes". In: *Current Opinion in Microbiology* 14.5 (2011). Antimicrobials/Genomics, pp. 615–623. ISSN: 1369-5274. DOI: https://doi.org/10.1016/j.mib.2011.07.027. URL: http://www.sciencedirect.com/science/article/pii/S1369527411001111.

[26] David A. Rasko et al. "The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates". In: *Journal of Bacteriology* 190.20 (2008), pp. 6881–6893. ISSN: 0021-9193. DOI: 10.1128/JB.00619-08. eprint: https://jb.asm.org/content/190/20/6881.full.pdf. URL: https://jb.asm.org/content/190/20/6881.

[27] Devashish Rath et al. "The CRISPR-Cas immune system: Biology, mechanisms and applications". In: *Biochimie* 117 (2015). Special Issue: Regulatory RNAs, pp. 119–128. ISSN: 0300-9084. DOI: https://doi.org/10.1016/j.biochi.2015.03.025. URL: http://www.sciencedirect.com/science/article/pii/S0300908415001042.

[28] Matt Ravenhall et al. "Inferring Horizontal Gene Transfer". In: *PLOS Computational Biology* 11.5 (May 2015), pp. 1–16. DOI: 10.1371/journal.pcbi.1004095. URL: https://doi.org/10.1371/journal.pcbi.1004095.

[29] L. Rouli et al. "The bacterial pangenome as a new tool for analysing pathogenic bacteria". In: *New Microbes and New Infections* 7 (2015), pp. 72–85. ISSN: 2052-2975. DOI: https://doi.org/10.1016/j.nmni.2015.06.005. URL: http://www.sciencedirect.com/science/article/pii/S2052297515000529.

[30] Sergey A. Shmakov et al. "The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes". In: *mBio* 8.5 (2017). Ed. by Michael S. Gilmore, Rotem Sorek, and Rodolphe Barrangou. DOI: 10.1128/mBio.01397-17. eprint: https://mbio.asm.org/content/8/5/e01397-17.full.pdf. URL: https://mbio.asm.org/content/8/5/e01397-17.

[31] Adi Stern et al. "Self-targeting by CRISPR: gene regulation or autoimmunity?" In: *Trends in Genetics* 26.8 (2010), pp. 335–340. ISSN: 0168-9525. DOI: https://doi.org/10.1016/j.tig.2010.05.008. URL: http://www.sciencedirect.com/science/article/pii/S0168952510001083.

[32] C. Than et al. "Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions". In: *J. Comput. Biol.* 14.4 (May 2007), pp. 517–535.

[33] C. J. von Wintersdorff et al. "Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer". In: *Front Microbiol* 7 (2016), p. 173.

[34] Bridget N. J. Watson, Raymond H. J. Staals, and Peter C. Fineran. "CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction". In: *mBio* 9.1 (2018). Ed. by Joseph Bondy-Denomy and Michael S. Gilmore. DOI: 10.1128/mBio.02406-17. eprint: https://mbio.asm.org/content/9/1/e02406-17.full.pdf. URL: https://mbio.asm.org/content/9/1/e02406-17.

[35] Sebastien Wielgoss et al. "Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load". In: *Proceedings of the National Academy of Sciences* 110.1 (2013), pp. 222–227. ISSN: 0027-8424. DOI: 10.1073/pnas.1219574110. eprint: http://www.pnas.org/content/110/1/222.full.pdf. URL: http://www.pnas.org/content/110/1/222.

[36] A. Zambelis, U. J. Dang, and G. B. Golding. "Effects of CRISPR-Cas System PResence On Lateral Gene Transfer Rates In Bacteria". 2015.

[37] Quan Zhang and Yuzhen Ye. "Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements". In: *BMC Bioinformatics* 18.1 (Feb. 2017), p. 92. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1512-4. URL: https://doi.org/10.1186/s12859-017-1512-4.

[38] Yan Zhang et al. "Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis". In: *Molecular Cell* 50.4 (2013), pp. 488–503. ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2013.05.001. URL: http://www.sciencedirect.com/science/article/pii/S109727651300364X.

[39] Olga Zhaxybayeva and W. Ford Doolittle. "Lateral gene transfer". In: *Current Biology* 21.7 (2011), R242–R246. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2011.01.045. URL: http://www.sciencedirect.com/science/article/pii/S0960982211001011.

# 6 Network Theory

Network theory is an extension of graph theory, a branch of mathematics concerning the properties of "graphs". Graph in this context refers to a set of nodes and a set of edges between those nodes, with edges typically representing some kind of relationship between those nodes[23]. Network theory focuses on modeling interactions using graphs and applying tools built to analyze graphs to gain an understanding of networks and how they function.

Consider a social networking site like Facebook, are people more likely to be friends with people who have a similar number of friends? Modeling this with a network, nodes represent users and an edge between nodes represent whether two users are Facebook friends. To answer the above question the assortativity of the network can be calculated. Assortativity is a measure of the network's nodes preference to form edges with more similar nodes[23]. Similarity here refers the difference in the number of edges connected to each node, *i.e.* the number of friends either user has. If a network has a large assortativity value, it means that similar nodes do connect to each other more often than different ones.[23]. Thus our question can be reformed through the lens of network theory as "Does a network constructed from Facebook user data have high assortativity?"

While the above example is a simple network, this model can be further extended through:

- Adding directions to edges (from one node to another)
- Adding weights to edges (often representing data about node interactions)
- Adding attributes to nodes themselves (binary,discrete or continuous)

An example of a directed, weighted biological network with low assortativity be a gene expression network (nodes:genes, edges:transcription level correlations) with few transcription factors which each modulates expression of multiple unrelated genes. For this project, nodes will represent OTUs and edges will represent genetic exchange between those OTUs, whereas in a normal phylogenies, edges only represent taxonomic relationships. Despite the complexity of HGT, network theory allows a flexible theoretic framework to analyze these interactions that are normally ignored by traditional phylogenetic methods.

# 7 Objectives

Using sequenced genomes, the goal of this project is to construct phylogenetic networks for all strains within sets of genera with and without CRISPR-Cas systems. Ultimately the goal of this project is to examine the relationship of HGT rates and the presence of CRISPR-Cas systems, using a network theoretic approach. The following sets of comparisons will contribute to the understanding of this relationship:

**Within Network Comparisons** For genera with strains containing CRISPR and Non-CRISPR species, comparing the network dynamics of those sets of nodes across genera will elucidate if CRISPR-Cas systems affect the HGT rates or the association patterns of individual OTUs.

**Between Network Comparisons** Next networks created from genera with no know CRISPR system containing strains (nc-networks) will be compared to mixed networks, containing strains both with and without CRISPR Systems. This will help understand whether the presence of CRISPR nodes can affect HGT network dynamics of OTUs other than themselves. A simple example may be that if mixed networks show more over all transfers across the network than nc-networks, CRISPR containing OTUs may be increasing HGT among closely related Non-CRISPR OTUs.

**Gene Indel Rates Vs. Network Statistics** Comparing insertion and deletion rates separately can help further specify what mechanisms may be responsible for trends observed in network statistic. If a mixed network is found to be density connected, but also shows an deletion bias, this may imply that most of the genes being transferred may not confer a fitness advantage.

# 8 Phylogenomic Networks

HGT is clearly an important factor in understanding evolution in prokaryotes. Since HGT has been found to be frequent throughout the prokaryotic tree of life, this has lead many to re-evaluate the concept of a 'tree of life', which by definition ignores these horizontal interactions[16].

## 8.1 Prokaryotic Net Of Life

In graph theory a tree is defined as a graph where there is only one path between every pair of nodes. In phylogenetics this implies there is only one path for genetic material to transfer between organisms, that path being vertical inheritance. As HGT demonstrates, this tree model is clearly an incomplete representation of genetic relationships between OTUs. Genetic material can be transfered outside of reproduction, allowing for multiple paths by which a single gene can be found in two OTUs (either inheritance, transfer or some combination of the two)[39]. This prompted the idea of a prokaryotic network of life (as opposed to a tree), with edges indicating both vertical and horizontal transfers of genetic material[16]. Edges can now connect OTUs to closely or distantly related OTUs, and even extinct ancestral OTUs normally found in phylogenetic trees.

## 8.2 Detection

While understanding that HGT is important and networks provide a useful theoretic framework to study it, constructing such networks is not trivial. Many different strategies have been developed to detect potential HGT events given a phylogenetic tree, with some potentially able to detect both recipients and donors[28]. There are two primary sets of methods

for detecting HGT.

**Parametric** These methods rely on investigating the sequence composition (GC%,Codon Bias, etc.) in genes and when they deviate from the genomic average. Average GC content has been found to vary significantly between some organisms, even by up to 30% in closely related organisms[28]. The same is true for codon bias, where codons are observed with different frequency in different bacteria, dependent on the expression levels of the tRNAs in those respective organisms[28, 17]. For example if *E. coli* contains more copies of a tRNA with the anti-codon TTA (Leucine) than CTC, genes will more likely encode the TTA codon to increase transcription efficiency[17]. If you see more TTA codons that CTC codons in a gene in *S. aureus*, assuming *S. aureus* has no leucin codon bias, one may be able to infer that the codon-biased gene was transferred horizontally[28]. Other metrics to consider are GC%, k-mer frequency or the presence of other features around the candidate gene, such as transposases or flanking sequences[28].

**Phylogenetic** These methods rely on recognizing discordance between gene trees and species trees. If a gene tree is found to have a significantly different topology from a species tree, this difference may be the result of an HGT event[32]. One can also compare the substructures of a gene trees and species trees (created by removing a set of edges leaving a set of sub-trees) to see if the substructures disagree[28]. Another strategy involves pruning (removing an edge to get 2 distinct trees) an internal branch and reattaching the subtrees at a different location. If the re-grafted tree has a better fit to the reference tree than the original, this may be indicative of an HGT event between the original node and the node the subtree was re-grafted to[28].

While HGTs can lead to these discordances, there are other series of evolutionary events than can produce the same results[32]. Events that may lead to false diagnosis of HGT are: incomplete lineage sorting, gene duplication followed by loss in one of the descendant lineages or homologous recombination[28, 32]. Strategies to account for these events, as well as account for uncertainty in your species or reference tree exist, but there still exist many cases with significant uncertainty[32].

It should also be noted that many of these methods require heuristic solutions, as they are computationally expensive, and sometimes even entirely intractable, which creates further uncertainty in the results obtained[28]. As an example, finding the minimum edit path between 2 trees (as in the re-grafting method) is NP-Hard, but the solution space can be limited by not considering pruning branches between consistent nodes[15, 28].

Generally phylogenetic methods are preferred for multiple reasons:

- Can make use of multiple genomes at once[28]
- Require explicit evolutionary models, which come with their own framework for hypothesis testing and model selection[28].
- HGT events identified by parametric methods are often found by phylogenetic methods as well[28].

- In recent years, the requirements of computing power and multiple well sequenced genomes for phylogenetic methods have become easier and easier to meet[28].

While detecting HGT events with high degrees of certainty is still difficult, much progress has been made in recent years,especially using phylogenetic methods[28].