UNIVERSITY of OTAGO

SAPERE AUDE

Te Whare Wānanga o Otāgo

NEW ZEALAND

UNIVERSITY OF OTAGO

MASTERS THESIS

The Twilight Zone of Nucleotide Homology

Author:
Stephanie MCGIMPSEY

Supervisor:
Paul GARDNER

*A thesis submitted in fulfilment of the requirements
for the degree of Master of Science*

in the

Department of Biochemistry
Univeristy of Otago

May 15, 2019

Department of Biochemistry

UNIVERISTY OF OTAGO

Abstract

Master of Science

The Twilight Zone of Nucleotide Homology

by Stephanie MCGIMPSEY

Homology search tools are important for inferring homology in the abundance of genomes currently sequenced. These tools utilise sequence similarity in order to assign a score between two sequences from which homology is inferred. The relationship between sequence similarity and homology can break down for certain levels of similarity. The zone of pairwise identity where a known pair of homologs has a 50% chance or less of being inferred as homologous based on the alignment score is called the twilight zone. The twilight zone for nucleotide homology has previously been calculated using databases that were small or contained bias. Therefore, the aim of this research was to calculate the twilight zone of nucleotide homology using a carefully designed database of homologous sequences. A database of core ncRNA and mRNA genes from a large range of genus representative bacteria was generated, from which sequence pairs were chosen. The database was used to calculate where the twilight zone of nucleotide homology was for four different types of alignment algorithms; `BLASTn`, `ggsearch`, `nhmmer` and `ssearch`. The effect of G+C content and sequence length on the location of the twilight zone was also examined. The twilight zone was shown to be between 38-50% pairwise identity for all alignment algorithms tested. Both sequence length and G+C content shift the twilight zone for all four alignment algorithms. This research has shown that between 38-50% pairwise identity homology should not be inferred based only on the alignment score, as there is a greater chance of incorrectly inferring homology than correctly inferring homology. Furthermore, the analyses have shown that a parametric approach to database design is required to further balance the database used for the twilight zone calculation.

Acknowledgements

I would like to thank Paul Gardner my thesis supervisor whose office door was always open whenever I ran into a trouble or had a question about my research. He allowed me to steer this research in my own direction, but always had suggestions of how to get back on track when things went awry.

I will be forever grateful to Kate Ryder from the Biochemistry Department for proof reading this thesis. I have learnt so many invaluable lessons about scientific writing and how to formulate a thesis from Kate that I will be able to utilise for the rest of my career.

I would also like to thank the Biochemistry Department, in particular the administration staff and the BIOC5F coordinator Lynette Brownfield, for helping with the logistics of preparing a masters thesis for examination and dealing with my general confusion about how things work here at the University of Otago.

A special shout out goes to my masters thesis committee, Sarah Diermeier and Xochitl Morgan. Their feedback and insights throughout the duration of the research helped me to narrow down on what the most important aspects to investigate were.

A huge thank you goes to the Gardner BinfLab for helping out with questions about coding syntax and methods suggestions. The mix of educational backgrounds in the group has been a valuable resource to expanding my knowledge base.

Thank you to the University of Otago for giving me a Masters Scholarship that allowed me to study my Masters degree without having to worry about finances.

Finally, words cannot express my gratitude to my family and friends for providing me support and encouragement throughout the years of study. In particular Dads 'its not the end of the world' speeches and Mums constant supply of hand knitted socks have helped get me through the year of research (and a Dunedin winter).

Contents

List of Figures

List of Tables

List of Abbreviations

16srRNA	**16S** **R**ibosomal **R**ibo**n**ucleic **A**cid
C+G	**C**ytosine **G**uanine
CI	**C**onfidence **I**nterval
CM	**C**ovarience **M**odel
CPU	**C**entral **P**rocessing **U**nit
DNA	**D**eoxyribo**n**ucleotic **A**cid
FPR	**F**alse **P**ositive **R**ate
HMM	**H**idden **M**arkov **M**odel
HSSP	**H**omology-derived **S**econdary **S**tructure **P**roteins
ID	**Id**entity
iTOL	**I**nteractive **T**ree **o**f **L**ife
mRNA	**M**essenger **R**ibo**n**ucleic **A**cid
ncRNA	**N**on-**C**oding **R**ibo**n**ucleic **A**cid
NCBI	**N**ational **C**enter for **B**iotechnology **I**nformation
ORF	**O**pen **R**eading **F**rame
PDB	**P**rotein **D**ata **B**ank
PID	**P**airwise **Id**entity
PPV	**P**ositive **P**redictive **V**alue
RNA	**R**ibo**n**ucleic **A**cid
SCOP	**S**tructural **C**lassification **O**f **P**roteins
SIMD	**S**ingle-**I**nstruction **M**ultiple-**D**ata
SSE2	**S**treaming **S**IMD **E**xtensions **2**
SSV	**S**ingle ungapped **S**egment **V**iterbi
tRNA	**T**ransfer **R**ibo**n**ucleic **A**cid

Chapter 1

Introduction

The invention of high-throughput sequencing has lead to an exponential increase in the number of sequenced genomes in the last 20 years (Fig. 1.1)(1, 2). The relative ease at which we can now access genetic information has coincided with an increase in research into understanding the complexity and composition of genomes, particularly using comparative methods. For large data sets, computational approaches are needed to effectively and quickly find features of interest (3). Pre-existing mathematical algorithms and heuristic techniques, particularly those relating to the field of graph theory, have been adapted to help with the processing, mapping and understanding of the genome (4–10). While these algorithms are often mathematically optimal, the biological sphere is full of exceptions and oddities that leave these algorithms often working less than perfectly (11). The majority of these algorithms are of non-deterministic polynomial time (NP) complexity which means in their pure form they can take an extremely long time to solve even the simplest of biological data sets (12). The field of bioinformatics tends to lead in the creation and adaption of algorithms using smart, biologically sensible heuristics that best map the biological processes occurring in the genome without losing the understanding of the underlying biology. This has lead to great advances in the ability to understand the relatedness of species or genes, understand the potential functions of genes without having to isolate the associated proteins or RNA and essentially 'time travel' to model evolution (13–16). This thesis explores the limitations of a subset of these software advances, homology search tools, and when you can use sequence alignments to infer homology between sequences.

1.1 Homology and similarity

Homology is a biological concept of great interest to biologists from many fields and is defined as having a common evolutionary origin between sequences, species or organisms (17–19). It is generally inferred by similarity (e.g. morphological traits, gene sequences), and this forms the basis for how homology is identified between gene sequences computationally (20). Homology is a binary question, things are

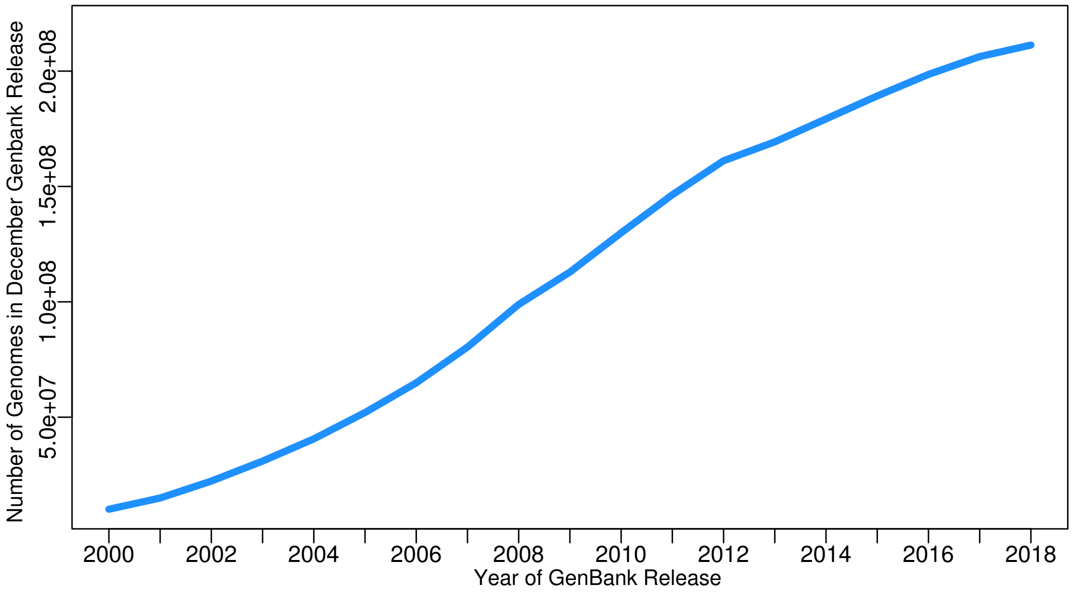


FIGURE 1.1: **Increase in the number of genomes available over the last 20 years:** Number of genomes in the December NCBI GenBank release each year graphed from data publicly available at https://www.ncbi.nlm.nih.gov/genbank/statistics/

either homologous or not, whereas similarity is a scale from nothing in common to everything in common. Therefore to superimpose the definition of homology to a gene on the basis of similarity requires us to levy limits of how much similarity is required before we can determine two things are homologous (*21*).

While homology between two sequences may elude to common ancestors, similar sequences are not always the result of common evolutionary origin. Other factors such as convergent evolution, low complexity sequences, sequence length or G+C content may cause high sequence similarity and none of these allow for inference of common ancestry (Fig. 1.2 A)(*22*, *23*). Random pairs of sequences have some level of similarity by chance purely due to utilising the same set of residues as each other (nucleotides or amino acids). On average two unrelated protein sequences of the same length will have 5% pairwise identity (the percentage of matching residues in the alignment, PID) if aligned without gaps whereas nucleotide sequences will have on average 25% PID (*19*). As we don't have access to most extinct ancestral genomes there is no easy way to check if genetic similarity is due to ancestry or other mitigating factors. Therefore the problem of assigning homology for a pair of sequences lies with how we determine similarity and at what levels of similarity we can be assured that homology exists (*19*). Homology determination is also complicated by the fact that homology may be divided into several sub types, and the different definitions of these may also cause issues when trying to infer homology between two species (*24*).

1.1.1 Parology, orthology and xenology

Homology is a catch all term that contains different sub-types that impact how genes can be used in alignments to infer homology. Orthology, parology and xenology are homology mechanisms with different biological interpretations (25). Orthologs are genes that diverged post speciation from a common ancestral gene (26). Paralogs are genes that are duplicates of an ancestral gene (26). Xenologs are homologous genes found in a species due to horizontal gene transfer (21). So within one genome it is entirely possible to have an ortholog, paralog and xenolog of the same gene all with different evolutionary origins. Widely used alignment algorithms lack the ability to differentiate between the types of homologs without the use of phylogenetic trees (27, 28). This leads to uncertainty as to whether an identified homolog is the ortholog that was intended, or if it is instead a paralog or xenolog.

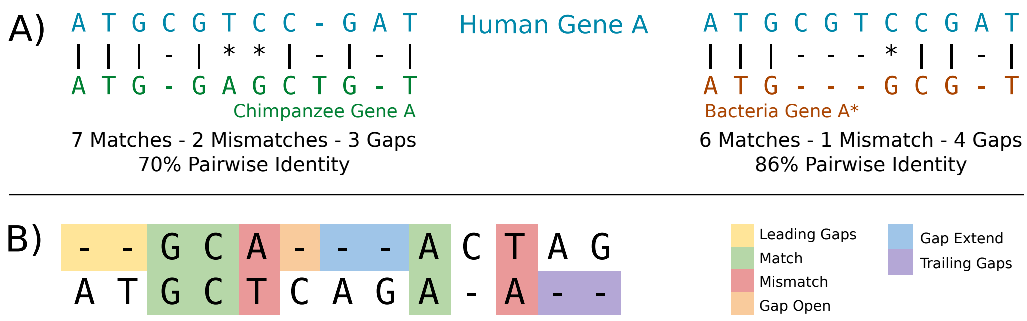


FIGURE 1.2: **Similarity versus homology and alignment scoring:** *A)* Two sequence alignments to Human Gene A highlighting how true homologs (Chimpanzee Gene A) can have similar similarity to non-homologous sequences (Bacteria Gene A*). *B)* A sequence alignment with different reward or penalties highlighted to illustrate how homology search algorithms score alignments based on sequence similarity.

1.1.2 Alignment and scoring

Currently the most common way to infer homology is to measure similarity between two sequences by aligning them using an algorithm to get a pairwise alignment (20). Each alignment algorithm calculates a score as a measure of how similar the two aligned sequences are and homology can be inferred using the score as evidence of significant similarity (11). Irrespective of the underlying mathematics of the algorithm, each alignment score depends on the location of matches, mismatches and gaps. Matches are where the two sequences have the same nucleotide (or amino acid for proteins) in the same place in the alignment and this is scored positively (Fig. 1.2 B). Mismatches are where there are two dissimilar nucleotides in the same place in the alignment (Fig. 1.2 B). These imply a substitution mutation has occurred in one of the sequences at this position and are often scored negatively as this makes the sequences less similar.

Gaps can be split into two categories; within the alignment and trailing or leading gaps (a.k.a. “end gaps”). Leading and trailing gaps are gaps at the very start or very end of the alignment (Fig. 1.2 B). They are often neither penalised or positively scored due to the fact that we could just be missing the start or end of one of the sequences due to problems with sequencing or assembling the genome. This helps prevent large penalties in the case of non-coding RNA (ncRNA) sequences as the exact start and end of the gene is unknown because they do not utilise mechanisms to define their boundaries like start and stop codons in protein coding sequences. Internal alignment gaps indicate indels (insertion-deletion events) (Fig. 1.2 B). These decrease the similarity between sequences so are also often scored negatively.

Some alignment algorithms distinguish these internal gaps into a further two categories; gap opening and gap extension. Gap opening means it is the very first gap in a sequence of gaps or a lone gap whereas gap extension means that a gap already precedes it. The biological reasoning behind this is that often insertions and deletions occur in one single evolutionary event (e.g. one insertion of multiple residues) so they need to be scored in such a way that long gaps are not overly penalised compared with short gaps when both could feasibly have occurred as one mutation event. These can be penalised differently or the same depending on the algorithm settings. The severity of penalisation or reward for each of the above cases is dependent on the parameters set by the algorithm and the algorithm type itself.

To determine true homologs from non-homologs requires the alignment score and statistical measures that allow for the significance of the score to be determined (20, 29). The alignment score is dependent on both the scoring scheme chosen and the underlying algorithm for the alignment tool.

1.2 Alignment algorithm types

Homology alignment tools are used in many different facets of biology due to their ability to determine potential function or the level of relatedness between sequences quickly and cheaply compared with traditional laboratory work (3). This means a variety of algorithms have been developed using different underlying mathematical principles to determine potential homologs (9, 30, 31). Alignment algorithms can be broken down into two groups, either by the way they align, or by the mathematical algorithm type.

1.2.1 Global and local alignments

There are two common ways to align two sequences; globally or locally (Table 1.3)(3, 32–34). Global alignments align sequences from start to finish whereas local alignments find highly conserved areas between the two sequences and don't necessarily align the whole sequence. Local alignments are better at aligning more diverged species that may only have functional domain conservation rather than full sequence conservation (34). The mathematical algorithms that underpin both search types are Needleman-Wunsch (global) and Smith-Waterman (local) (4, 5). Both algorithms utilise a recursively built matrix of alignment scores for each residue in the two sequences as well as traceback methods to determine the alignment from how the matrix scores were calculated (33). The recursive matrix is built using a set of formulae that differs between Needleman-Wunsch and Smith-Waterman.

Needleman-Wunsch

For Needleman-Wunsch the initial top left cell of the matrix (corresponds to a gap-gap cell) is initialised to 0 and the rest of the first row (sequence 1 aligned to all gap columns) and column (sequence 2 aligned to all gap columns) are filled with decreasing negative multiples of the gap penalty (5, 33). After the initialisation of the first row and column each cell in the matrix is recursively calculated based on its three neighbouring cells that preceed it to determine which cell, added to either a match score or gap penalty, will be the maximum value. The largest value is added to the cell and an arrow added pointing to which cell of the three neighbours the score was calculated from (33). If multiple values are the same an arbitrary choice is made although both potential origin cells could be kept if a graph structure rather than a matrix was used. The whole matrix is filled similarly and once complete the bottom right cell will be the score for the optimal alignment. Starting at the cell, the traceback arrows are used to find that alignment from the bottom right cell to the top left cell. This will give the optimal alignment of both sequences to each other. At certain points in the traceback there may have been two different ways to get the same score and thus slight variations to give the same optimal alignment score are possible.

Smith-Waterman

Smith-Waterman differs in the recursion by filling cells in the matrix with a zero if the recursive score (calculated from the three preceding cells) would otherwise be negative (4, 33, 34). The traceback is also different as the highest score found in the matrix is the starting point for tracing back the alignment rather then the bottom right matrix cell. This allows the alignment to end and start anywhere in the matrix. The traceback ends when a cell of value 0 is encountered so the alignment also does not have to reach the top left hand corner of the matrix either. This allows alignments of parts of the sequences rather then the whole sequence length. There may be multiple partial alignments that have the same score and these alignments can all be found from the matrix.

Both of these methods are computationally costly (although Smith-Waterman is more computationally intensive as multiple different alignments can be formed for a sequence pair) and therefore some alignment algorithms have estimations of the methods (using heuristics) rather then the full computation.

1.2.2 Underlying mathematics of homology search tools

Homology search tools can be broken down by algorithm type into those that are quick database search heuristic methods and those that are dynamic methods (use Smith-Waterman or Needleman-Wunsch in their entirety) (Fig 1.3). Dynamic algorithms are guaranteed to find the optimal alignment for the two sequences for a given scoring scheme, although the alignment given may not be the most biologically relevant, whereas heuristic approaches may only find good alignments (not the mathematically most optimal) (*33*, *35*). The loss in accuracy is due to approximations of dynamic algorithms being made using heuristic techniques that are done to increase the speed of alignment compared with dynamic algorithms (*33*). Four commonly used homology search algorithms for nucleotide sequences, `BLASTn`, `nhmmer`, `ssearch` and `ggsearch`, all represent different combinations of local, global, dynamic and heuristic approaches (Tab. B.1) (*36*–*38*).

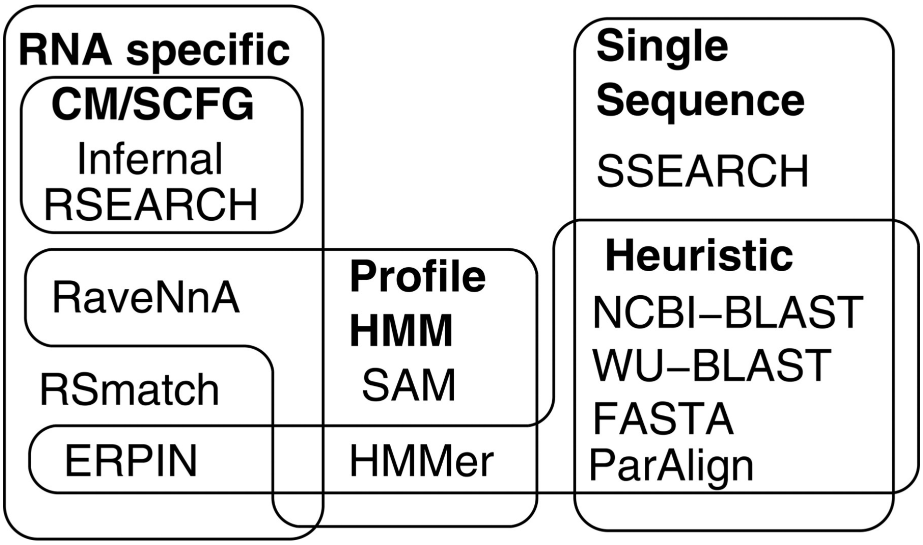
`ggearch` **and** `ssearch`

`ggsearch` implements Needleman-Wunsch to score each alignment between the query and the database as described above (*30*, *37*, *39*). `ssearch` uses Smith-Waterman which is accelerated by striped Single-Instruction Multiple-Data (SIMD) referred to as Streaming SIMD Extensions 2 (SSE2) (*4*, *37*, *40*). This basically performs the Smith-Waterman operation as described above but on multiple parts of the matrix in parallel to speed up computation of all possible local alignments for two sequences (*40*).

`BLASTn`

The `BLASTn` algorithm divides the query sequence up into nucleotide sections (words) of 11 nucleotides (default parameter setting) (*3*, *9*, *38*). Similar words to those found in the query (e.g. have one or two different nucleotides) are also computed and aligned to the word they originate from. Similar words that are above a score threshold are included in the word set. Matches between words and database sequences are then extended in either direction until the alignment score drops below a set

An overview of homology search methods.



Eva K. Freyhult et al. Genome Res. 2007;17:117-125

FIGURE 1.3: **Homology search algorithms overview:**[a] An overview of homology search methods. A Venn diagram illustrating an overview of the methods used in this study. Different methods are classified as heuristic, single sequence, profile HMM, stochastic context-free grammar (SCFG), and/or RNA specific.

[a] Adapted from Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA, Freyholt *et al.*. 2007, published by Genome Research, Cold Spring Harbor Laboratory Press. Permissible by CC-BY license.

threshold. A heuristic approximation of the Smith-Waterman algorithm is used to create gaped alignments. Each alignment is given a score based on the scoring scheme. The default score for a match is 2, mismatch is -3, gap open is 5 and gap extend is 2 (*41*). To access the significance of alignments an E-value is calculated for all potential alignments. An E-value for a sequence match is the number of sequences in a database expected to have a similar score as the match by chance based on the database size (Eqn. 1.1) (*34*). Those that have an E-value above the given threshold are included in the final output. The use of words to find regions of similarity makes `BLASTn` a fast alignment tool, as not every combination of sequences has an alignment score calculated. However BLAST is not guaranteed to find the optimal alignment (*34*, *42*). Despite this limitation, it will find good alignments between sequence pairs and the E-value statistic provides the user with more information about the significance of the alignment in addition to the alignment score.

$$
\begin{gathered}
\textbf{E-value} = Kmne^{-\lambda S} \\
K\text{: Scale factor for the search space, } m\text{: Length of the query sequence} \\
n\text{: Number of sequences in the database, } \lambda\text{: Scale factor for the scoring system} \\
S\text{: Alignment score}
\end{gathered}
\tag{1.1}
$$

`nhmmer`

`nhmmer` utilises probabilities and profile hidden Markov models (HMM) in a six step process to narrow the database sequences to statistically significant matches (36, 43). A HMM is a statistical model that describes the potential evolutionary outcomes for each residue in a sequence based on the probabilities of each evolutionary event, calculated from a multiple sequence alignment for a gene (10, 44). There are 3 hidden states which are the evolutionary events that aren't directly observable from the sequence; match, insert and deletion (Fig. 1.4) (44). This emulates evolution in that we see the residues in a sequence but we don't see the evolutionary events (e.g. deletion) that caused that sequence to exist. This allows for more flexibility in scoring at certain residues in a sequence when aligning a sequence to an HMM as its not just a residue to residue comparison. For each residue in the sequence there is a probability that it is an insert, deletion or a match that is calculated from the model based on what residue is at the position in the sequence. This allows for variable regions to be less heavily penalised if there is high variation and highly conserved regions to be heavily penalised if there is variation.

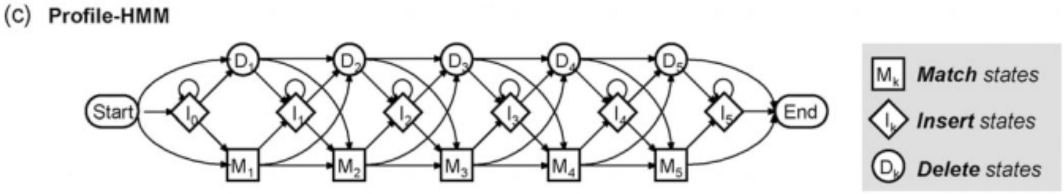


FIGURE 1.4: **Profile HMM diagram:** Diagram of a profile HMM showing the hidden states of an HMM (insert, match and delete) and the transitions possible between them. [a]

[a] Reproduced fromHidden Markov Models and their Applications in Biological Sequence Analysis, Yoon 2009 under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium.

An alignment of sequences by `nhmmer` to a profile HMMs produces a bit score as the measure of similarity of a database sequence to the HMM. A bit score is the log two odds ratio of the probability of the profile HMM given the data and the probability of the null model given the data (Eqn. 1.2)(45). The null model (a single state HMM emitting nucleotides at equal frequencies) is used to calculate the probability that the sequence is not homologous to the profile HMM (36, 43, 46). Each step in the `nhmmer` pipeline calculates bit scores using this method.

$$\textbf{Bit score} = log_2 \frac{P(model|data)}{P(null|data)} \tag{1.2}$$

For nhmmer the single ungapped segment Viterbi (SSV) heuristic filtering step is used to align all database sequences without gaps in a local alignment to a match state only HMM (36). Striped Single-Instruction Multiple-Data (SMID) is used to speed up the algorithm (40). High scoring segments of the sequence are extracted in order to prevent entire chromosomes or contigs being further processed (if the database contains genomes) (46). A p-value threshold is set for the distribution of bit scores based on comparison with a pre-calculated null distribution of random sequences, and sequences that pass this threshold pass on to the next filter. The p-value threshold prevents the majority of non-homologs from passing onto the next step. The score calculated during SSV filtering is adjusted based on estimated sequence composition bias in the potential homolog. Any sequence with an adjusted score above the same p value threshold as the SSV filtering step is passed through the Viterbi filter. A gap permissible alignment to the profile HMM is done using the Viterbi algorithm which is accelerated using SIMD vector implementation. The Viterbi filter calculates the most probable hidden state passes for the observable sequence, which means it can find the sequence of match-insert-delete states that would emit the observed nucleotide sequence with highest probability for the given profile HMM (33). A new p-value threshold is applied to the distribution of scores produced and those that are above it are scored with the full Forward parsing algorithm.

The Forward algorithm calculates the total probability that the database sequence was generated by the HMM by summing all the probabilities of each possible state path that could emit the observed sequence (33). Essentially it finds all the combinations of match-insert-delete (paths) that could emit the sequence and adds the probabilities of each path together. It is a recursive algorithm as the probability of each state is dependent on the states that came before. The Forward algorithm is very similar to Viterbi except it uses all the possible paths rather than the maximal scoring one to score the database sequence. Only if the score passes one final p value cut off is the sequence also aligned using the Backward algorithm to the profile HMM. The Backward algorithm calculates the probability of each state (match, insert, delete) for each residue in the sequence which is akin to the posterior probability of the states except it is calculated in reverse recursive (i.e. starts at the end of the sequence) (33). This probability is used to calculate the bit score for each local alignment.

1.2.3 Defining homology from a pairwise alignment

A pairwise alignment is where two sequences are aligned using an alignment algorithm and allows for direct comparison between the sequences (34). The alignment similarity score is used to infer homology between sequences and is influenced by the alignment algorithm used, along with the scoring scheme. Most algorithms calculate a bit score; a measure of sequence similarity independent of alignment length and database size (36, 39, 41). Similar sequences have a higher bit score and scores above a set threshold (determined by the user) are inferred as homologous.

Its also worth noting that observable changes between two sequences will not be the only sequence mutations that have occurred over time (29). Reverse mutations and multiple mutations at the same residue are obscured by comparing two sequences at one point in time (29). The process of comparing similarity to determine relatedness lacks the ability to account for the potential of these 'invisible' mutations. This means the divergence of two sequences can be underestimated (29). The reverse can also happen where two sequences have diverged in their sequence similarity to the extent they have not maintained enough similarly for their homology to be detectable using sequence similarity so the relationship between the two sequences can't be identified(29). Sequence similarity is one of the best measures available for determining homology, alongside structural similarity, but it is fundamentally flawed especially when trying to determine homology for distant relatives (20).

Correctly inferring homology based on sequence similarity in an alignment is important for many different processes such as annotation of genes and taxonomy classification. Errors in homology inference can cause many different problems that can have an effect on the wider scientific community. Incorrect assignment of function can cause a gene to be assigned a potential function it does not have but it can also cause the erroneous annotation to propagate to other genome annotations or sequence databases (e.g. glyoxalase I function annotated for a soybean gene but the sequence had no match to the Structure Function Linkage Database HMM for glyoxalase I) (47–49). Incorrect assignment of taxonomy based on alignments in metagenomics studies can cause errors in identifying which species are present in an environment (e.g. *Yersinia pestis* (aka the plague) as a 'normal' part of the New York subway microbiome) (50, 51). Therefore determining the homology between pairwise alignments relies on having checks in place to prevent homology being inferred when there is not enough evidence for it.

Pairwise alignments can be grouped into four categories using the bit score and the structure of the gene (52). Actual homologs tend to have similar sequence and same structure whereas sequences likely to be called as false homologs have similar sequence but dissimilar structure. Sequences that are dissimilar but have similar structure are presumed cases of convergent evolution. The final case is sequence pairs considered unrelated which have dissimilar sequence and dissimilar structure.

The ability to differentiate between true and false homologs is important for genetics research however alignments with low similarity scores (potentially distant homologs or unrelated sequences) can be indistinguishable (19, 20, 29). As sequence similarity decreases between aligned sequence pairs the ability to detect true homologs trends downward and the number of false homologs increases (52). At some level of sequence similarity the number of false homologs exceeds the number of true homologs. The zone at which this occurs has been labelled the twilight zone ((11, 29, 32, 52–54).

1.3 The twilight zone

The term twilight zone was first coined in relation to pairwise alignments (of nucleotides or amino acid sequences) in 1986 by Russell F. Doolittle (29). The twilight zone for pairwise alignments is the range of sequence similarity for which the relationship between similarity and homology breaks down to the extent its not possible to accurately infer homology using sequence similarity. The reason inferring homology in the twilight zone is difficult is it is characterised by a massive increase in non-homologous sequence pairs relative to true homologs paris for the same sequence similarity (52). Therefore the alignment bit score, or some other score of similarity, is no longer enough evidence to infer homology without supplementary data as it is too difficult to distinguish true from false homologs (11). Non-homologous sequences with similar pairwise similarity to true homologs occur due to sequence artifacts such as repeats, sequence composition (e.g. high G+C content), short sequence length or convergent evolution.

The twilight zone occurs for a different range of sequence similarity for nucleotide and proteins alignments due to the different alphabets utilised. Proteins have 21 amino acids (including selenocysteine) in their alphabet and nucleotide sequences have four bases. Nucleotide alignments are expected to have a higher twilight zone than amino acid alignment due to the smaller alphabet.

There are two different types of twilight zones for alignments. One twilight zone identifies the loss of sensitivity in alignment accuracy of two sequences (11, 32). This is where the alignment of two sequences is compared for accuracy to a trusted reference alignment. The other type of twilight zone determines the loss in sensitivity to inferring homology between two sequences (29, 52, 54). The shared similarity of aligned sequences is used to determine if two sequences are homologous or not and the twilight zone begins where the probability of correctly inferring homology is 50% or less. The second type of twilight zone will be addressed in this thesis.

1.3.1 Protein alignment twilight zones

The initial concept of the twilight zone for protein alignments was developed by Doolittle in 1986 and has since been expanded upon by Rost and Thompson *et al.* in 1999 (52). Rost determined the twilight zone for inferring homology whereas

Thompson *et al.* determined the twilight zone for alignment accuracy (32, 52). Since 1999 little has been done to determine if the twilight zone of homology has changed as more algorithms have been written or algorithms have been updated. Unpublished research from 2012 that was divulged via a first person interview with Sean Eddy, one of the leading experts in homology alignment algorithms, is the only available comparison to Rost's work in 1999 (54).

Doolittle's twilight zone is between 15-25% and is a 'rule of thumb' for sequences of 100 amino acid length or greater (29). The 100 amino acids caveat is due to the greater probability of short sequences having the same amino acid sequence by chance and therefore the level of confidence in short alignments is not as easy to give general advice for (19, 29). No quantitative experiment was done to determine this zone, it was merely a guideline to follow for researchers to have more confidence in their ability to determine relatedness from the alignment. At this point in biological history, alignments were a new source of information, and the importance of this was not understood as well as it is today (19). The outcome of this book chapter was that it is difficult to determine if a sequence pair in the twilight zone is truly homologous based on just the alignment and these pairs requires further computational evidence to determine their homology. Doolittle ultimately concluded that when determining if sequence pairs that lie in the twilight zone are homologous, these pairs require further computational evidence to determine their homology (29). Despite this initial concept being suggested by Doolittle, it was not experimentally calculated until 1999 by Rost (52).

Rost showed that the twilight zone of homology for protein alignments was between 20-35% pairwise identity (52). It should be noted that the main aim of this paper was not to determine the twilight zone however the author did calculate characteristics that identify the location of it. The twilight zone was highlighted by a 5000 fold increase in aligned protein pairs that were falsely considered homologous (similarity in sequence composition but not in structure) meaning false positives outweighed true positives for this zone of pairwise sequence identity (Fig. 1.5 A) . This resulted in making the identification of true homologs in this zone a difficult task(52). The database Rost utilized had 1.7 million pairwise alignments of 792 unique proteins with known structure aligned to 5,646 amino acid sequences from the Protein Data Bank (PDB) (56). The three alignment algorithms used included a Smith-Waterman based full dynamic program MaxHom and two quick database search methods BLASTP (heuristic Smith-Waterman) and PSI-BLAST (heuristic Smith-Waterman) (9, 57). Blast was relatively new at the time and was an important addition to the alignment algorithm family due to its speed. Rost showed that BLASTP or PSI-BLAST alignment similarity scores were less accurate than those of MaxHom at separating true and false homologs for the same coverage (also know as precision or positive predictive value (PPV)) (52).

In the same year that Rost calculated the twilight zone for homology, Thompson

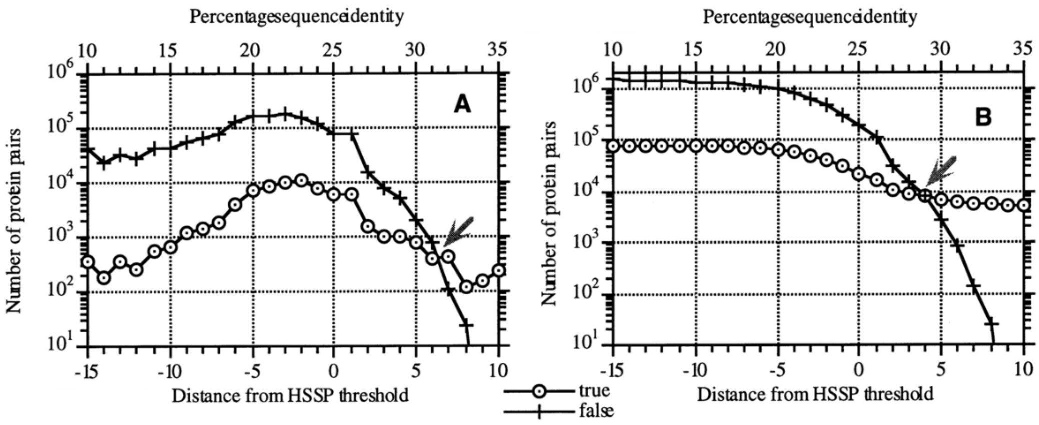


FIGURE 1.5: **Rosts twilight zone of protein homology**[a] Explosion of structurally dissimilar pairs in the twilight zone. Numbers of true (pairs with similar structure) and of false positives (pairs with no similar structure) plotted versus the distance to the HSSP-curve (55), i.e. the horizontal axes give the distance from the threshold defined in eqn 1 (numbers refer to the parameter n in eqn 1). The levels of pairwise sequence identity corresponding to the distance were shown on top. (A) Number of pairs observed at any distance (logarithmic scale). (B) Cumulative number of pairs observed (logarithmic scale). For example, at a threshold corresponding to about 32% sequence identity for long alignments, the numbers of true and false positives were equal (arrow in A); at about 29% even the cumulative numbers of true and false positives were equal (arrow in B). Note: numbers of true negatives and false negatives result from the cumulative sums left of the threshold; percentages of true and false positives given in Figure 5.

identified the twilight zone for alignment accuracy to be (10-20% pairwise identity) (32). The accuracy of an alignment does not necessarily report about the homology of the sequence pairs. It only shows that the alignment generated by the algorithm matches with a certain percentage to that of a trusted alignment. Whilst incorrect alignments of sequence pairs could cause them to have lower PID than expected (due to not have the maximal number of matched residues), the work of Thompson *et al.* does not apply directly to the scope of research of this thesis.

The twilight zone calculated by Eddy in 2012 shows that the lower threshold of the twilight zone is now as low as 11% (Fig. 1.6) (54). HMMER was first published in 1995 so was available when Rost did his research into the twilight zone but he did not include the algorithm in his experiment (52). The database set up for Eddy's experiment is very different then Rost. A pairmark database was created where a single protein domain target is embedded in a large non-homologous sequence similar to the profmark benchmark methodology (43, 58, 59). This was done to create a database that mimics databases that contain whole contigs, chromosomes or

[a] Figure adapted with permission from Oxford Univeristy Press, Protein Engineering, Design and Selection;Twilight zone of protein sequence alignments, Rost 1999. Copyright Feb-1999. Reproduced with Permission 4518510088473.

genomes, rather then individual sequences for each gene. The database created had 10,691 query sequences searched against 142,322 real sequence targets and 200,000 negative targets (shuffled sequences) (54). Four different alignment algorithms were tested, HMMER 3.0, PSI-BLAST, BLASTP and Smith-Waterman, however two different scoring schemes where tested for Smith-Waterman (Fig. 1.6). The fraction of true homologs detected for each PID was calculated for fixed specificity, the true negative rate (the proportion of true negatives found out of all negative targets present). The lowest twilight zone of all the alignment algorithms was HMMER 3.0 with 11% (Fig. 1.6). Eddy is the author of the suite of programs in HMMER 3.0. This plus the lack of peer review for this research would suggest that this calculated zone should be used with caution (31, 46).

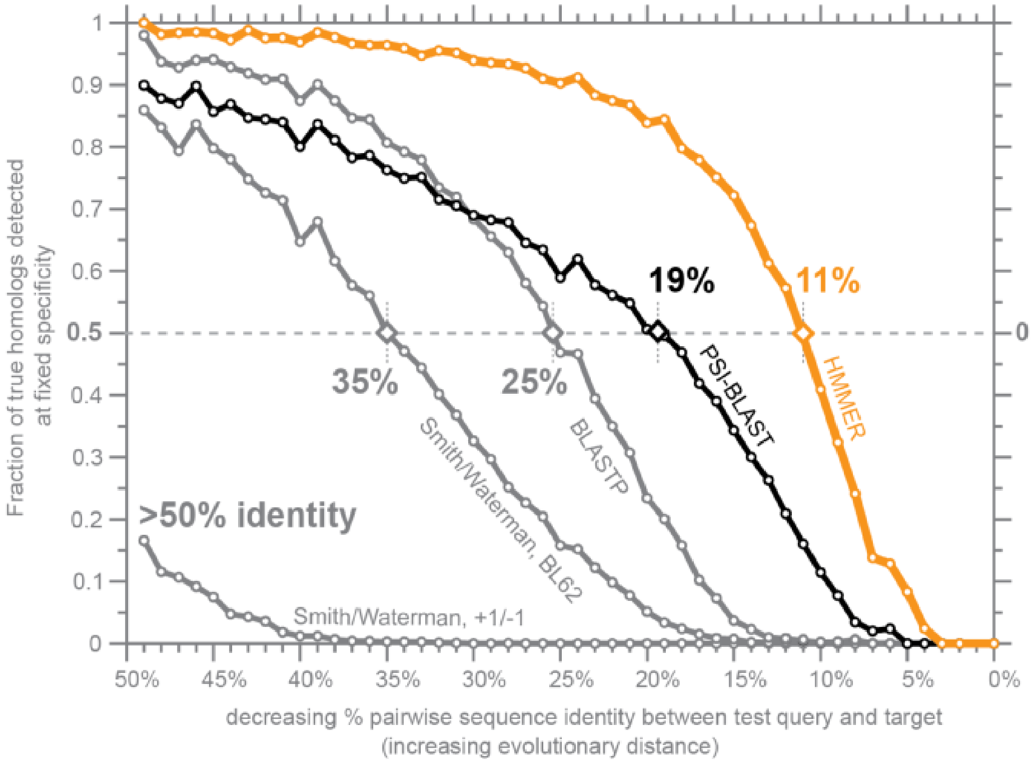


FIGURE 1.6: **Eddys twilight zone of protein homology:** Twilight zone analyses using a pairmark database approach to calculate the fraction of true homologs detected at fixed specificity (true negative rate). Only sequence pairs between 0-50% were tested. This research has not been pair reviewed.

Whilst there have been advancements in research into the amino acid alignment twilight zone and the improvements can be seen in our ability to detect homology within it, less research has been done on the nucleotide alignment twilight zone.

1.3.2 Nucleotide alignment twilight zones

The twilight zone for nucleotide alignment accuracy was first calculated by Gardner *et al.* in 2005 whereas the twilight zone for homology was calculated in 2010 (*11*, *53*). There was a delay in calculating the twilight zones for nucleotide alignments despite the ability to sequence DNA being available since 1975 when Sanger *et al.* described the Sanger sequencing method (*60*). Potentially the difficulty in detecting ncRNA due to their absence of distinctive start and end points meant there was a lack of sequence data for determining the twilight zone until covariance models (CMs) were created to find ncRNAs in genomes (*61*).

As a part of a study of structural ncRNAs that were used to benchmark multiple sequence alignment programs the twilight zone for nucleotide alignment accuracy was calculated (*11*). It was between 50-60% which is much higher then the 10-20% calculated by Thompson *et al.* 1999 for protein alignments (*32*). This zone represents pairwise identities where little to no structural information can be gathered from structurally aligned homologous ncRNAs using alignment algorithms alone. The authors used a set of four RNA alignments that were seed alignments for CMs from Rfam v5.0 as their reference alignments similar to that of Thompson *et al.* 1999. This data set was later published for open access and named the BRaliBase after the amino acid sequence BAliBASE database (*11*, *62*, *63*). The much higher twilight zone for ncRNA was attributed to the difference in alphabet size (4 nucleotides versus 21 amino acids) resulting in a higher probability of gaining some sequence patterns by chance, especially if they lack complexity (*19*, *29*).

The database utilised for this calculation of the twilight zone was identified to have a flaw, nicknamed the BRailiBase dent, where even the most advanced structural RNA alignment tools seem to perform weakly at best in the 60-40% pairwise identity range (*64*). This was due to tRNAs being at least half the sequence pairs or more for PID values below 40% (*64*). tRNAs are structurally well conserved but lack sequence similarity (*65*). All the algorithms used in this benchmark align ncRNAs using structure and sequence so even for low sequence pairwise identity values alignment algorithms are able to align tRNA to a high level of accuracy (compared with other ncRNA genes for that PID) based mostly on structure. Other ncRNA's don't tend to be as structurally well conserved as tRNA for PID values below 60%. Within the twilight zone (40-60%) there are still alignments from other genes that make up at least 50% or more of the data for those PID values. Therefore these alignments, which won't be as accurate as tRNA alignments due to less structural conservation will help balance out the unusual nature of the tRNAs structural conservation. Below the twilight zone there are very few other ncRNA sequence pairs so alignment accuracy is based on only tRNA alignments unlike for higher PIDs. This means after the twilight zone there was a perceived increase in the accuracy of alignments causing a dent like trend in the alignment accuracy of all alignment algorithms (Fig. 1.7.

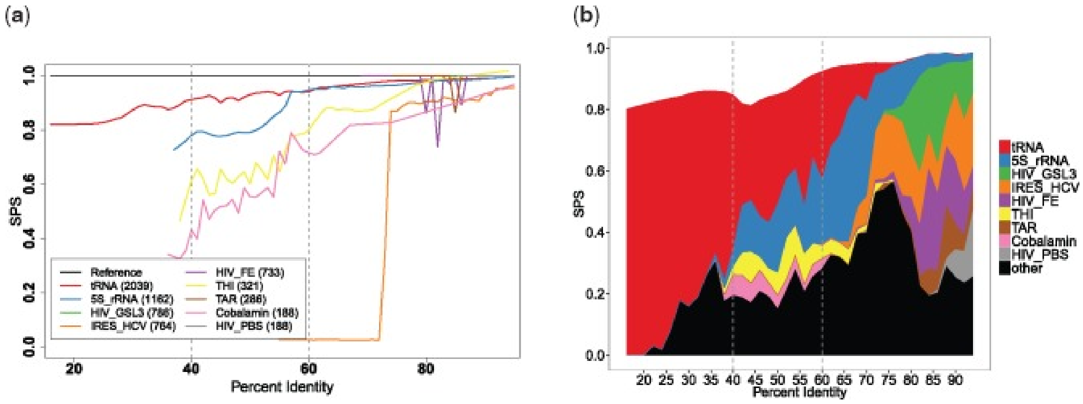


FIGURE 1.7: **The BRaliBase Dent**[a] The two plots show 9 of 36 RNA families with at least 180 alignments. (A) Familywise performance of LocARNA. The family names in the legend are further accompanied by the total number of alignments for each family in brackets. (B) Each family's share of LocARNA's SPS (after local regression) per sequence identity. The remaining families with <180 alignments are grouped into 'other'. These graphs show the proportion of genes in the enhanced BRaliBase database (62). Only tRNA, 5s rRNA, U5 spliceosomal RNA and Group II introns were used in the Gardner *et al.* twilight zone calculation.

[a] Figure adapted with permission from Oxford Univeristy Press, Briefings in Bioinformatics;The BRaliBase dent—a tale of benchmark design and interpretation, Löwes, Benedikt; Chauve, Cedric 2016. Copyright March-2016. Reproduced with Permission 4518541063092.

The dent was considered to be the twilight zone as it was thought something odd occurred for this zone of PID that alignment algorithms could not cope with when in fact it was just due to a biased data set below the dent.

Whilst this twilight zone calculation is not comparable to the research done in this thesis it stands as a good example of how one gene being over represented in a data set can cause an unusual result.

The next attempt at calculating the twilight zone came before the BRalibase dent was identified, however the authors used a set of non-identical (<100% PID) structural RNA molecules from the Protein Database (PDB) (53). The main aim of the research was to quantify the relationship between sequence and structure conservation in RNA hence why they used RNA structures in PDB found using xray crystallography, nuclear magnetic resonance imagining (NMR)or electron microscopy. Crystallisation of an RNA often requires the sequence to have sticky ends added, non-essential sections removed or stem loops added to help make the RNA structural more ordered to obtain crystals (66, 67). This is because RNA molecules are very flexible, especially in large loop or linker regions. This can mean the RNA is found in a conformation that isn't necessarily the most common structure for the molecule when compared with structures determined in solution via NMR (68). These RNA structures from PDB were used alongside sequence similarity to determine homology between RNAs (53).

Their analyses allowed them to quantify a twilight zone for ncRNA homology for Infernal v1.0 using single sequence targets searched against covariance models (CMs) generated from the remaining sequences in the data set. They used the Infernal E-value as a threshold to separate true and false homologs based on structure and sequence conservation. They found that homologous pairs of ncRNA sequences will diverge structurally when there is a significant reduction in their sequence PID. As the relationship between sequence and structural conservation becomes less correlated this makes it hard to determine if the amount of structural conservation is significant enough to infer homology for low PID values (<60%). Distant homologs had sequence PIDs between 30-60% so the relationship between structure and sequence conservation for these sequences is more variable.

Therefore between 30-60% PID inferring homology based on the relationship between sequence and structure was harder, thus indicating the twilight zone. The methodology of this paper mimics that of Rost *et al.* 1999 where sequences of known structure are used to identify how well alignment algorithms can be used to infer homology from sequence alone (*52*). The authors only use one alignment algorithm and as Gardner *et al.* 2005 showed, different alignment algorithms, with different underlying mathematical approaches, have different twilight zones so basing the twilight zone off one algorithm is a narrow approach to the problem (*11*). Their database is biased towards RNA structures that have been crystallised or imaged and are found in PDB. This does not necessarily represent all RNA molecules known to date, nor does it represent a wide taxonomic range of bacteria for each RNA molecule.

1.4 Proposed research and summary

The overarching theme with previous attempts at calculating the twilight zones for nucleotide sequences is that both have database flaws that make them non-representative of all RNA's (*11*, *53*). They either focus on a small subset of ncRNA where there is structural data available or that there is a trusted reference alignment that has been curated. Their focus on ncRNA has resulted in the omission of a large section of RNA sequences; messenger RNAs (mRNA). Whilst the homology of mRNA cannot be easily confirmed by structural conservation like that of ncRNA, it can be confirmed by the conservation of amino acid sequence and protein structure (*19*, *52*).

In order to improve on the previous attempt to calculate the twilight zone of nucleotide homology by Capriotti *et al.* 2010 three aims were set. The first was to compile a data set comprised of a representative sample of core mRNA and ncRNA genes from a wide phylogenetic range of bacteria genera that was as unbiased as possible. The idea was that this data set would be representative of nucleotide alignments as a whole so the twilight zone can be applied in any nucleotide pairwise alignment.

The next aim was to calculate the twilight zone for nucleotide homology, utilizing the data set generated, by using different alignment algorithms that represents all of homology search tool space. The four algorithms chosen were `ggsearch` (global and dynamic), `ssearch` (local and dynamic), `BLASTn` (local and heuristic) and `nhmmer` (local and heuristic + dynamic (with HMMs)). Each algorithm is widely used and has over 2000 citations between their original paper and their newest update (Tab. B.1) (*20*).

The final aim was to observe any differences for older versions of alignment algorithms compared to the new versions when it came to the ability to extract true homologs in the twilight zone. This was to see if improvements had been made over time in alignment algorithms ability to align pairs of sequences in such a way that we can more easily separate true and false positives.

Chapter 2

Methods

An overview of the experimental design has been summarised in figure (Fig. 2.1) and all code, alongside a detailed computational pipeline, can be found at the Github link in table (Appendix. C.1).

2.1 Database curation

2.1.1 Genome download

All bacterial genomes (101,133) available on the 27th of November 2017 were downloaded over a 6 day period from the NCBI RefSeq database using their supplied ftp instructions (Fig. 2.2) (*69*). Contig identifiers for each genome were checked to determine if they were unique for genomes containing multiple contigs. This was done by comparing each contig identifier in a genome file and removing any genomes with contigs that had identical headers. Any genome that had one or more redundant contig labels was removed from the set of genomes to prevent incorrectly calling homologs. This is because algorithms used in this research use the contig identifier and nucleotide coordinates within the contig to identify genes. The resultant genomes are subsequently referred to as the genome set.

2.1.2 Taxonomy ranks

For each genome in the genome set, taxonomy ranks were obtained from the NCBI taxonomy database. Two publicly available data files, accession2taxid and taxdump, were used by taxonomizr v0.2.2, an existing R code package, to get taxonomy rank data for each NCBI accession ID. The NCBI accession ID is converted to a NCBI tax ID using the accession2taxid file and the taxID is used to find the taxonomy information in taxdump. A custom piece of code (language Perl v5.26.0) was written to create a tidy tabular output of the taxonomy ranks for each accession (Appendix C.1). The same code made master files for each taxonomy rank (e.g. class, genus) where each name in a rank has a list of corresponding genome accessions stored

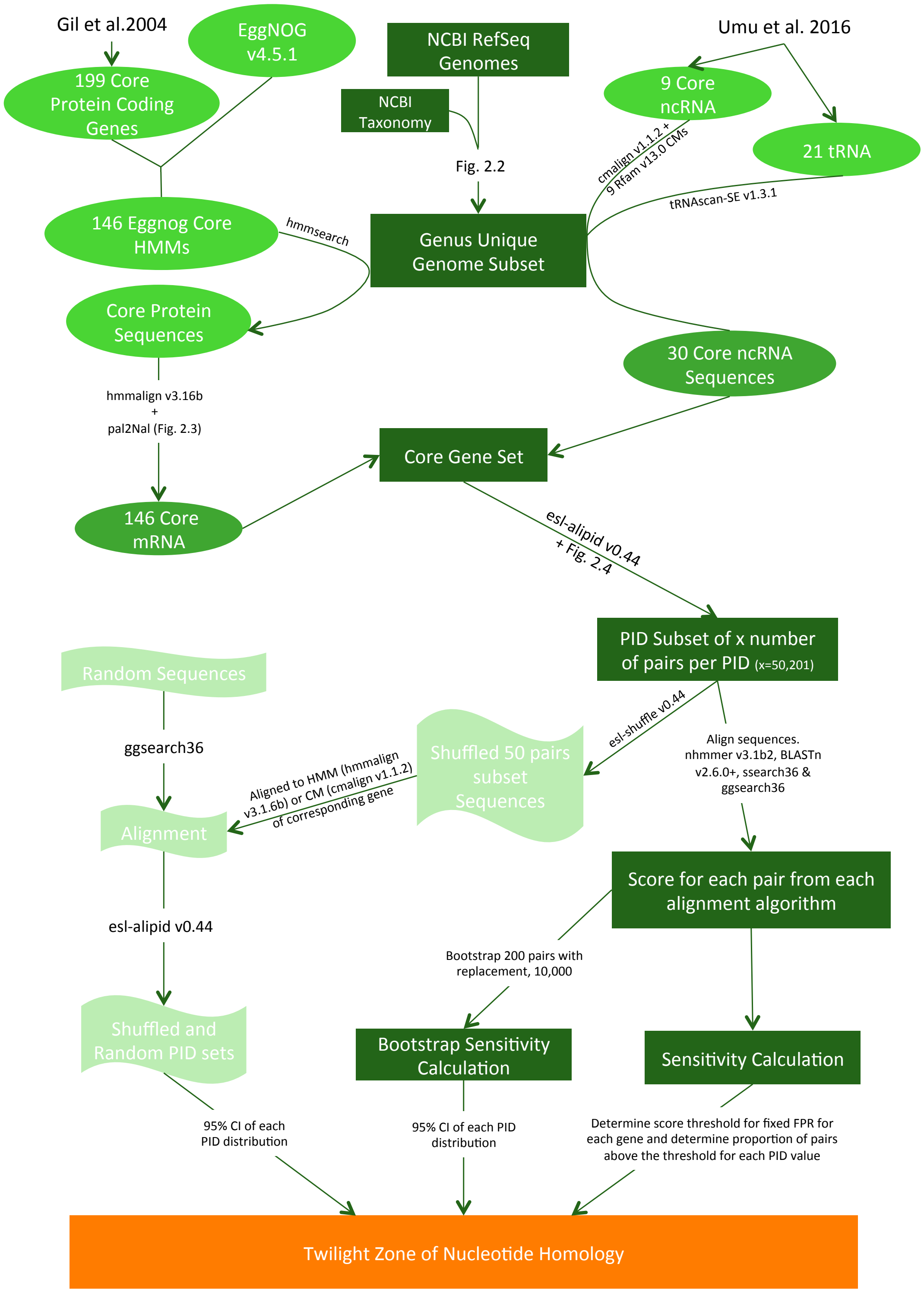

Gil et al.2004
EggNOG v4.5.1
NCBI RefSeq Genomes
Umu et al. 2016
199 Core Protein Coding Genes
NCBI Taxonomy
9 Core ncRNA
21 tRNA
Fig. 2.2
cmalign v1.1.2 + 9 Rfam v13.0 CMs
tRNAscan-SE v1.3.1
146 Eggnog Core HMMs
hmmsearch
Genus Unique Genome Subset
Core Protein Sequences
30 Core ncRNA Sequences
hmmalign v3.16b + pal2Nal (Fig. 2.3)
Core Gene Set
146 Core mRNA
esl-alipid v0.44 + Fig. 2.4
PID Subset of x number of pairs per PID (x=50,201)
Random Sequences
esl-shuffle v0.44
ggsearch36
Shuffled 50 pairs subset Sequences
Align sequences. nhmmer v3.1b2, BLASTn v2.6.0+, ssearch36 & ggsearch36
Aligned to HMM (hmmalign v3.1.6b) or CM (cmalign v1.1.2) of corresponding gene
Alignment
Score for each pair from each alignment algorithm
esl-alipid v0.44
Bootstrap 200 pairs with replacement, 10,000
Shuffled and Random PID sets
Bootstrap Sensitivity Calculation
Sensitivity Calculation
95% CI of each PID distribution
95% CI of each PID distribution
Determine score threshold for fixed FPR for each gene and determine proportion of pairs above the threshold for each PID value
Twilight Zone of Nucleotide Homology

FIGURE 2.1: **Methods overview:** Simplified overview of computational methods detailing the experimental design that lead to the calculation of the twilight zone for nucleotide homology. A more detailed pipeline is available on Github (Appendix C.1)

within the file. A NCBI taxonomy tree was produced from this data to show the phylogenetic range of all genomes based on their phylum. Any species with an unknown phylum classification was excluded from the tree. Each phylum was further split into classes and each class and phylum was labeled with the number of species it contained. The tree file was created by transforming the tidy tabular output into a Newick format tree with a polytomy for each phylum and set of classes. The tree is a pictorial representative of the genome set (Fig. 3.3). The tree was visualised using the Interactive Tree of Life (iTOL) online tool and annotation templates they provide (70).

2.1.3 Gene and genome selection

Genes that span a wide phylogenetic range were required so there would be varying amounts of divergence between homologs to gain a wide range of pairwise identity (PID) values for each gene (20). Pairwise identity is defined as the percentage of similar residues between a pair of aligned sequences (Eqn. 2.2) (71). Core bacterial genes, both ncRNA and mRNA, were selected from previously published 'essential' gene sets found in scientific literature (72, 73). Nine core ncRNA genes covariance models (CMs) were downloaded from Rfam v13.0 and all 144,498 available EggNOG v4.5.1 profile hidden markov models (HMMs) were downloaded on the 9th of April 2018 (74, 75).

HMMs are used to identify homologs in a diverse set of genomes due to their ability to 'forgive' divergence in unessential regions of genes which is often present in pairs of species that diverged a long period of time ago (10). A CM is a probabilistic model similar to an HMM except it also encodes secondary structure information as well as nucleotide sequence information based on a structural multiple sequence alignment for a gene (61). These models are used to find homologs for ncRNA genes rather than a single sequence as often ncRNA is structurally more conserved than the nucleotide sequence is. Therefore sequence diverged homologs can still be found as long as they retain structural features of importance to the gene. Due to the large and diverse taxonomic set of bacterial genomes used in this research CMs were required to identify ncRNA gene sequences in all bacterial species.

The 9 ncRNA genes make up the ncRNA core gene set which excludes tRNAs as the Rfam model cannot distinguish between tRNA types (e.g. Alanine vs Valine). The tRNA core gene set consists of 21 tRNAs predicted by tRNAscan-SE v1.3.1 that incorporate the 21 proteinogenic amino acids into proteins utilized by most bacterial phylum (76–78).

mRNA core gene set

The set of EggNOG profile HMMs contained genes that were not in the core gene set, as well as multiple profile HMMs annotated to the same core gene. The set was reduced to one representative profile HMM per core gene by using 199 essential protein genes for *Escherichia coli* str. K-12 substr. MDS42 and *Bacillus subtilis* str. 199 to select the best profile HMM for each core gene (73). These genomes were used as they are from different diverged phylum and both genomes have well characterised core genes (79). All sequences were downloaded from NCBI. These sequences were searched against all 144,498 EggNOG profile HMMs using hmmsearch v3.1b2 with an E-value threshold of 1e-10 . Profile HMMs that matched both *E.coli* and *B.subtilis* core genes were extracted and the bit scores compared by taking the proportion of the smallest bit score relative to the sum of both bit scores (Eqn. 2.1). If the proportion is 1 then the model only matched either *E.coli* or *B.subtilis*. A profile HMM that has the same bit score for both *E.coli* and *B.subtilis* sequences for a gene would have be scored as 50% using this relative measure.

If the proportion was between 30% to 70% the EggNOG profile HMM was selected. If a core gene had multiple profile HMMs in this range the model with the closest proportion to 50% was chosen. These proportion thresholds were used to ensure the profile HMM was not heavily biased to finding sequences in only one part of bacterial taxonomy distribution as the two genomes used to find the core gene profile HMMs were from phylum known to have structural and sequence divergence in proteins (80–82). Bias in finding sequences can happen if the alignment that the profile HMM is calculated from does not include sequences for that gene from a diverse range of bacterial species. The range is chosen to exclude models that aren't close to a 50:50 ratio of bit scores. EggNOG profile HMMs were used as they model the full protein sequence and not just protein domains like other profile HMMs such as Pfam (83). The downside is that they are not hand curated models like Rfam and Pfam nor is there one model per gene (15, 16). This is why the above process was required to gain a set of full length profile HMMs for each core gene.

This set of 146 profile HMMs is called the mRNA core gene set.

$$\textbf{Bit score proportion for } \boldsymbol{E.coli} \textbf{ and } \boldsymbol{B.subtilis} = \frac{min\{b_{E.coli}, b_{B.subtilis}\}}{b_{E.coli} + b_{B.subtilis}} \cdot 100 \tag{2.1}$$

$$\begin{aligned} b_{E.coli} &= \text{ the bit score for the gene from } E.coli \\ b_{B.subtilis} &= \text{ the bit score for the gene from } B.subtilis \end{aligned}$$

Genus representative genome set

A balanced and diverse taxonomic set of genomes was required to represent all sequenced bacteria. A combination of ncRNA and mRNA genes were used to select a representative genome for each genus. cmsearch v1.1.2 with default settings was used to annotate 16s rRNA in the genome set with the corresponding Rfam CM (RF0177) (*84*). Sequences that were shorter than 1000 or longer than 2000 nucleotides in length and had a bit score of 563.95 or more were ranked by bit score to find the top bit score match for each genome. 16s rRNA was used as it is thought to be present in all bacterial genomes and therefore, if it was absent from a genome in the set, it would be indicative of issues with the assembly or sequence quality (*85*). The bit score threshold was chosen by calculating the distribution of bit scores for 16s rRNA to find the means and standard deviations of the bimodal distribution (Fig. A.1). A bit score cut off of 563.95 was chosen as it is two standard deviations from the mean of the lower distribution. This method assumed both distributions were normally distributed but by inspection they are skewed (Fig. A.1). The cut off is therefore a rough estimate to try and remove the majority of (false) sequences modelled by the lower bit score mode distribution. As seen in the graphical representation of the bit score distribution for 16srRNA the cut off removes the bulk of the lower distribution sequences (Fig. A.1). The lower distribution generally contains noise or fragment matches to the Rfam CM (*86*, *87*). As the two distributions overlap it isn't possible to determine a cut off that divides the true homologs from other sequences however this cut off removes the majority of sequence that are partial or sequence composition matches. The length cut offs of 1000 to 2000 were chosen to select sequences of approximately the expected length (around 1500 nucleotides) and remove very short or very long matches that may have made the bit score threshold. The average length of 16srRNA was 1542 nucleotides long and the seed alignment for the Rfam 16srRNA model had sequences that range in length from 1424 to 1629 nucleotides long (*74*).

One final test was used to check genome quality in terms of whether multiple protein coding genes required for an essential function, such as DNA replication, could be located (*88*). Six DNA replication genes are in the core gene set and the profile HMMs were searched against a 6 frame translation of every genome using hmmsearch v3.1b2 with E-value threshold 1 (*89*). The high E-value threshold was selected after a previous run with E-value of 1e-10 caused no sequences to be found for more diverged species. One of these species, *Deinococcus ficus*, was searched individually and all six DNA replication genes were found so the sequences were present in the genome. A re-run of the original search with E-value of 1 saw all six DNA replication genes for *D.ficus* found.

This occurred due to the E-value thresholds dependence on the database size. A high bit score is considered less significant for a larger database than a smaller database

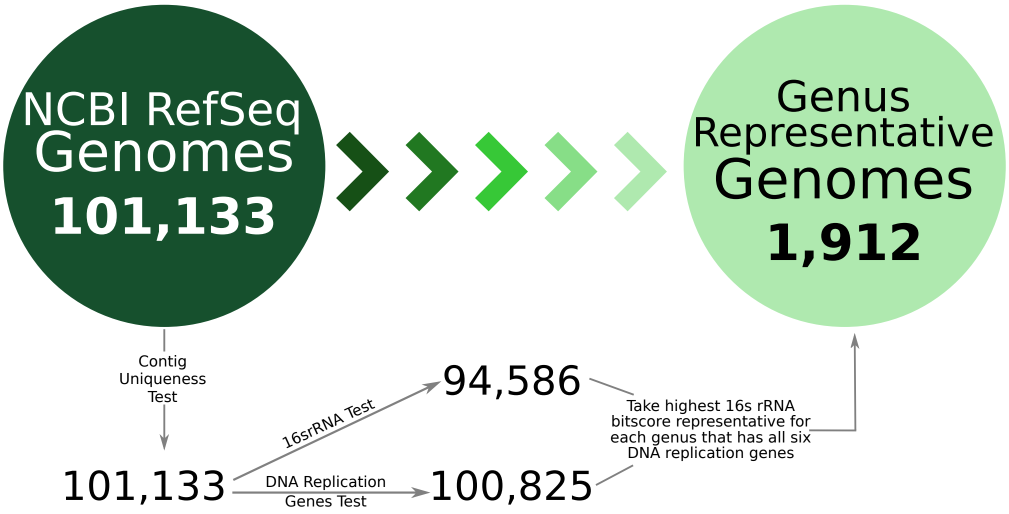

NCBI RefSeq Genomes
101,133
Genus Representative Genomes
1,912
Contig Uniqueness Test
101,133
16srRNA Test
94,586
DNA Replication Genes Test
100,825
Take highest 16s rRNA bitscore representative for each genus that has all six DNA replication genes

FIGURE 2.2: **Selection of a genus representative genome set:** Schematic demonstrating the steps taken to reduce the available set of bacterial genomes from RefSeq down to a much smaller set of good quality genus representatives

as the larger database has more sequences that could produce high bit scores purely by chance (*20*). Therefore it is easier to find more diverged homologous sequences when using an E-value threshold if the database is smaller as their lower bit scores will be considered more significant. There were 10,002,411,364 open reading frames (ORF) in the search database of length greater than 20 amino acids for 101,133 NCBI genomes. Genomes where the 6 DNA replication and 16srRNA genes were found were grouped based on their genus. The genome with the highest scoring 16srRNA gene sequence to the Rfam CM was selected to represent each genus (Fig. 2.2). There were 1,912 unique genera available (any with genus of NA or Candidatus were excluded from analyses as both denote undetermined taxonomic status so their true genus classification is unknown). This set of genomes is called the genus representative set (Tab. C.3 Link A).

Genus representative core gene set

mRNA and ncRNA homologs were found using sequence similarity to corresponding probabilistic models (Tab. B.3 & B.2). CMs contain information about consensus secondary structure alongside nucleotide sequence so the ncRNA sequences for each gene are found using structural and sequence similarity (*61*). HMMs do not include structure information as the structure of proteins is hard to predict from a sequence, so it is not as easy to encode structural conservation into a probabilistic model as it is for ncRNA. Therefore the protein sequences in the data set are found using sequence similarity to the HMM. The models used in the core gene set consist of 146 mRNAs, 9 ncRNAs and 21 tRNAs (Tab. B.3, B.2). These genes were all extracted from the genus representative genome set via three different methods.

For the mRNA core gene set each genus representative genome was translated in 6 frames using esl-translate v0.44 (Easel 0.44). This step is required due to EggNOG profile HMMs only modelling the amino acid sequence for each gene rather than the nucleotide sequence. The next release of HMMER would circumvent this problem as a translated search option will be added to hmmsearch (*54*). hmmsearch v3.1b2 was used to annotate sequences for each gene in the mRNA core gene set for each genome using the corresponding EggNOG profile HMM. Sequences were only reported if their bit score was more than a third of the maximum bit score possible for each profile HMM and only the highest bit score match per gene for each genome was selected. The bit score cut off was chosen after visual inspection of the distributions of bit scores for each model which are all bimodal (*86*). The smaller bit score mode distribution for each gene can usually be attributed to sequence fragments and other unrelated sequences therefore these need to be removed from the sequence set. Allprotein and ncRNA models had a smaller bit score mode distribution that contained non-homologous sequences with bit scores smaller then one third the maximum bit score. This value was therefore used to exclude the sequences in the smaller bit score mode distribution.(Fig. A.1). Rather than set the cut offs by hand for all 146 genes the more automated 1/3 max bit score approach was applied as it roughly approximated where each cut off would manually be set. The maximum bit score for each model was calculated by omitting the consensus sequence for each profile HMM using hmmemit v3.1b2 and aligning it back to the profile HMM using hmmalign v3.1b2. Default parameters were used for both algorithms. The mRNA core gene sequences found were aligned to their corresponding profile HMM using hmmalign v3.1b2 to obtain a protein sequence alignment to be used to calculate amino acid sequence PID.

For the ncRNAs, each of the 9 Rfam CMs was searched against the genomes in the genus representative set using cmsearch v1.1.2 with a bit score threshold of one third the maximum bit score possible for each model. For consistency the same bit score cut off was used for ncRNA as mRNA despite the small number of models making hand curation of thresholds more reasonable (Fig. A.1). The highest bit score match for each genome was selected and these sequences were aligned to their corresponding CM using cmalign v1.1.2 to obtain an alignment to be used to calculate PID.

tRNAscan-SE v1.3.1 was used to annotate the 21 tRNAs from the tRNA core gene set for each genus representative genome. tRNAscan-SE v1.3.1 utilises its own internal CMs. Default settings were used for tRNAscan-SE v1.3.1. For each of the 21 core tRNAs the highest bit score match was selected for each genome. These sequences were aligned in their individual tRNA groups to the Rfam general tRNA CM using cmalign v1.1.2 to obtain an alignment to calculate PID from.

Each sequence in the ncRNA, tRNA and mRNA (amino acid sequences) alignment was further quality checked for length to remove sequences that cannot be confidently determined as homologous. Anything greater than 2/3 or less than 1/3 the median length of sequences in the alignment was discarded using esl-alimanip (Easel 0.44). These thresholds were chosen to remove sequences that had large indels or were truncated as these are harder to confidently determine the true homology status of. As these are core genes required for day to day function it is unlikely that shorter or longer versions will be functional genes.

Due to the large number of mRNA profile HMMs in the core gene set, any gene that had sequences for less than 900 genus representative genomes were removed from the mRNA core gene set. This left 127 mRNA core genes with a representative sequence for at least half the genomes. The remaining mRNA amino acid sequences were then reverse translated in frame of their amino acid alignment to a nucleotide alignment using pal2nal v14 (Fig. 2.3) (*90*). The codons to be used for each amino acid in the reverse translation are taken from the nucleotide coordinates in the contig associated with the open reading frame that the protein sequence was obtained from. This resulted in an alignment of the nucleotide sequences that is in the same frame as the protein alignment.

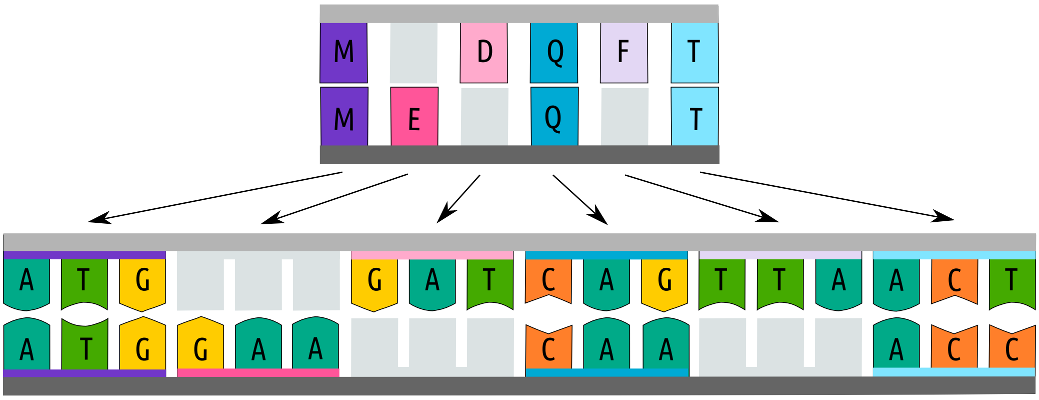


FIGURE 2.3: **Amino acid to nucleotide translation in frame:** The amino acid alignment is reverse translated to a nucleotide alignment keeping the codons in frame of the amino acid alignment. The codons used are from the nucleotide ORFs that correspond to the protein sequences.

This was done as the amino acid sequence is the structured, functional product of the mRNA and is the level at which natural selection acts. Therefore the nucleotide alignment should reflect the conservation of amino acids in order to be biologically more informative of evolutionary events (*19*, *20*, *29*, *90*).

The core gene set consists of 157 gene families from mRNA, ncRNA and tRNA that have representative sequences from genomes in the genus representative set. In total there are 275,394 sequences in this set comprised of 225,632 mRNA sequences, 37,768 tRNA sequences and 11,994 ncRNA sequences.

2.1.4 Pairwise identity and subset selection

Pairwise identity (PID) is a measure of similarity independent of the alignment bit score that is easy to calculate from an alignment. It was used as it is an easily calculated measure of evolutionary distance between two homologous sequences and is strongly correlated to more complex phylogenetic distance measures (e.g. F84 DNA distance) (Pearson correlation coefficient of -0.995) (Fig. A.3) (*20*, *91*). It was also used as the measure of similarity in previous twilight zone calculations for both protein and nucleotide sequences (*11*, *29*, *32*, *52*, *53*). PID can be calculated four different ways using different denominators; shortest sequence length, number of aligned residues including gaps, number of aligned residues excluding gaps or the mean length of both sequences (*71*). The research described in this thesis uses the most commonly found PID calculation for twilight zone analyses which divides by the shortest sequence length (Eqn. 2.2) (*11*, *19*, *29*, *32*, *52*, *53*). The core gene set alignments were used to calculate PID, as a percentage, using esl-alipid (Easel 0.44). All PIDs were rounded to the nearest 1%. There were 243,558,970 sequence pairs in total.

$$\textbf{Pairwise sequence Identity} = PID = \frac{\sum m_{ij}}{n_i}, \text{ for } n_i \leq n_j$$
$$n_i, n_j \text{ are the number of residues in sequence i or j respectively}$$
$$m_{ij} \text{ are the number of matches between pairwise aligned sequences i and j} \tag{2.2}$$

For each PID value (from 1-100%) a subset of 201 aligned sequence pairs was selected where available. This means the maximum number of pairs for this subset was 20,000. For each PID no more than 20% of the pairs comprised of one gene and all 21 tRNAs were treated as one gene (Fig. 2.4). 0% PID values were excluded from further analysis. The pairs of sequences selected are called the 201 pairs subset.

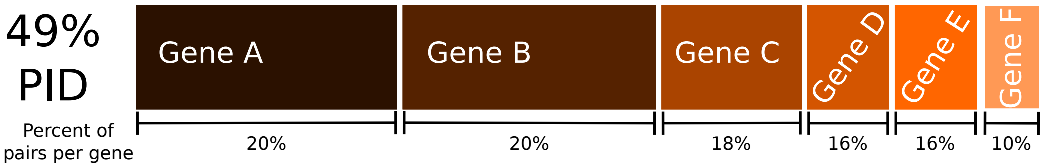


FIGURE 2.4: **PID value breakdown of pairs:** Example of the breakdown of pairs for a PID value (49%). Each gene makes up 20% or less of the pairs for the PID value and there are multiple different genes represented.

Additional PID subset

Between 68% and 100% of the 201 pairs subset, tRNA sequences were found to be over represented (≥75% of pairs) in a subsequent analysis of gene types. This resulted from a code specifier failing to count the 21 different tRNA families as a single group, and instead they were treated as 21 unique groups. As each gene, or group, can be up to 20% of the pairs per PID value this allowed tRNA to saturate the pairs for higher PID values. This is the previously reported problem that this experimental design was trying to avoid with our approach (64). As this problem was found late in the period of research it was unable to be rectified fully due to the time some algorithms take to run (Tab. 3.4). A second smaller subset of 50 pairs per PID was created and tRNAs were excluded from all PID values. This set of pairs is called the 50 pairs subset

2.1.5 Shuffled genes

In a database the majority of sequences are non-homologous to the query, so a negative control set of non-homologous genes was required to mimic true homology search space. Each gene in the core gene set was shuffled using esl-shuffle (Easel 0.44) to create a set of sequences that are known to be non-homologous but have similar sequence composition to the core gene set. ncRNA and tRNA genes were di-nucleotide shuffled whilst mRNA genes were codon preserving shuffled.

The dinucleotide and codon preserving non-homologous controls are one of the more conservative controls that could be used (72). They are a more challenging control set then mono-order Markov-model based sets and this reduces the risk of falsely claiming a high sensitivity when the real sensitivity is much lower (92–94) These sequences make up the bulk of the search space for each alignment algorithm to be tested. This set of shuffled sequences is called the non-homologous database.

A smaller non-homologous database was also created that included two shuffled versions of each of the sequences in the 50 pairs subset. This was done to decrease the amount of compute time it took to align sequences to the large non-homologous database. This smaller set of shuffled sequences is referred to as the 50 pairs non-homologous database.

Two further shuffled variations of each gene in the 50 pairs per PID subset were generated as a control experiment to determine the distribution of PIDs for sequences that are not homologous but preserve some higher order sequence features. The shuffled mRNA nucleotide sequences were translated in frame 1 to amino acid sequences. The real mRNA sequences do not include the stop codon at the end of each sequence (due being obtained by reverse translation from amino acid sequences) so shuffling these sequences did not cause premature stops within the amino acid sequence. Each set of shuffled sequences (amino acid or nucleotide) were aligned to the corresponding EggNOG profile HMM or Rfam CM for the gene they were derived from. The mRNA amino acid sequences were then reverse translated in frame the same way as for the PID subsets. The PID was calculated as above. This set is called the shuffled gene control set.

2.1.6 G+C and length simulation

Guanine-cytosine (G+C) content and sequence length can affect the ability to infer homology from an alignment. The repetitive nature of G+C rich regions and short sequences may result in two sequences having a similar bit score to true homologs purely by chance (20, 29). To determine the effect of G+C content and sequence length on the distribution of PID values for random sequences a set of sequences with fixed G+C and length requirements was randomly generated. There were equal numbers of 28% and 74% G+C content sequences (minimum and maximum G+C content for the genomes in the core gene set) with a length distribution that mimics that of the core gene set.

All sequence pairs were aligned one pair at a time using `ggsearch36` and their PID calculated based off that alignment using esl-alipid (Easel 0.44). Only sequence pairs where both sequences were between 1/3 and 5/3 of the length of the other sequence in the pair were included in the final PID distributions. The mode, minimum and maximum values of each distribution of PID's was calculated to determine how sequence features affect PID. This set is called the random sequence set.

2.1.7 G+C and length labelling

In order to explore how G+C content and length affects homology prediction, each sequence in the database was labelled based on its G+C content and length. The G+C content and length of each sequence was calculated using the sequence statistics from esl-seqstat (Easel 0.44). Sequences with a maximum G+C content of 40% were classed as low G+C content and labeled as G0. Sequences between 40% and 60% G+C content were considered medium G+C content and labelled G1. Sequences with a G+C content of 60% and above were considered to have high G+C content and labeled as G2. Similarly sequences were labelled as short (L0) if their length was less than 100 nucleotides, medium (L1) if their length was 100 to 999 nucleotides long and long (L2) if their length was 1000 nucleotides or longer. The shuffled gene control and random sequence sets were also labelled based on the G+C content and length.

The G+C content of the 1912 genus representative genomes (based on the median G+C content of the core genes found for that genome) ranged from 28% to 74% so the groups G0-G2 roughly divide this range into thirds.

The length groups were decided based on previous research as well as the range of lengths in the core gene set. Capriotti *et al.* 2010 found that the E-value threshold required to separate true and false homologs of length 100 or more was the same but sequences of shorter length required more stringent, lower E-value thresholds (*53*). E-values can be used to determine the level of acceptable background noise for determining significant sequence matches. In general sequences shorter than 100 nucleotides long require a more significant score in order to be included above the E-value threshold as E-value and score are inversely related. Therefore, 100 nucleotides was chosen as the threshold for sequences to be grouped as short sequences. The other threshold, 1000 nucleotides, was chosen as it was between the median and upper quartile of the sequence length distribution for the core gene sequences.

2.2 Twilight Zone

2.2.1 Homology search

Four commonly used homology search algorithms (`nhmmer v3.1b2`, `ssearch36`, `ggsearch36` and `BLASTn v2.6.0+`) were evaluated to determine how well they could distinguish true homologs from non-homologous sequences for each PID value (Appendix C.1 & C.2). All sequences in the PID subset being analysed (201 pairs or 50 pairs) from the same gene were grouped together, regardless of PID, and each gene was aligned independent of the others. The set of sequences for a gene along with the corresponding shuffled database make up the homology search space for each alignment algorithm. Each query sequence (one sequence from each pair of sequences in the gene) was aligned to this database and a bit score calculated (Fig. 2.5). All alignment algorithms parameters were set to ensure a consistent scoring scheme for match (+4), mismatch (-5), gap open (-10) and gap extension (-6) were used except for `nhmmer` which uses a probabilistic scoring scheme so the default scoring scheme was used (*36*, *95*). Parameters were selected such that as many matches in the database as possible were reported to ensure bit scores for shuffled sequences as well as true homologs were reported for each gene (Tab. C.2). In order to set the similarity threshold for determining homology a certain number of shuffled sequences need to have been reported.

Older versions of software

All alignment algorithms used in the twilight zone calculation have been available since the 90's except nhmmer which was a new addition to the HMMER suite of tools in 2013 (*9*, *30*, *36*). The oldest online version for each algorithm was located and tested for the ability to function in a Linux Ubuntu 17.10 environment. `ssearch`

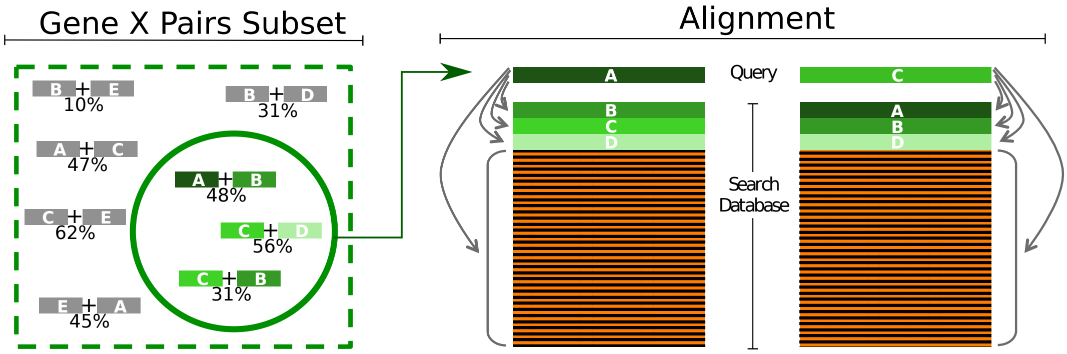


FIGURE 2.5: **Alignment procedure for each gene:** The pairs selected for a gene are collated together regardless of the PID value. One sequence for each pair is chosen as the query sequence. Each query sequence for the gene, in this case two (A and C), is aligned (using each of the different alignment algorithms) to the rest of the sequences for the gene as well as the shuffled database. The bit score of each query-database sequence match is recorded for each alignment algorithm.

v34.26.5 from April 2007 was the only software version that was still supported and could be installed. The same homology search procedure and algorithm parameters were used for ssearch34 as ssearch36 (Tab. C.2).

Iterative nhmmer

nhmmer v3.1b2 is unique compared to the other algorithms as it makes a profile HMM from the query sequences to search the database. The use of a single query sequence does not allow for the full possibility of a profile HMM to be utilised. This is due to the ability of profile HMMs to include information about the probability of indels, matches and mismatches at each position when a profile HMM is computed from more than one sequence. To test if nhmmer v3.1b2 would be more accurate at determining homologs in the twilight zone if given more evolutionary information about the query sequence, an iterative approach was designed.

The first nhmmer alignment iteration used a single sequence query to search against the database with an E-value threshold (0.1 for 201 pairs, 1e-10 for 50 pairs) to only include sequences that were a good match to the sequence, and therefore likely to be true homologs. The alignment of sequences above the E-value threshold was used to build a profile HMM using hmmbuild v3.1b2. The newly created profile HMM was then used as the query for the second iteration of nhmmer v3.1b2 to search against the database again but this time using a multi-sequence generated profile HMM. The same E-value threshold was used on the results from the second iteration to determine the set of sequences to be used to build the second profile HMM using hmmbuild v3.1b2. The second generation profile HMM was used for the final nhmmer v3.1b2 search of the database. The parameter choice for filtering and the reporting threshold was the same as the single sequence nhmmer v3.1b2 homology search

(Tab. C.2). The output of the third `nhmmer v3.1b2` homology search was taken as the final result.

The E-value threshold for the 50 pairs subset was chosen to be smaller then the 201 pairs subset. This is because evaluation of the sequences from the 201 pairs subset used to create profile HMMs for each iteration of `nhmmer v3.1b2` homology searching showed a very small number of non-homologous sequences were included in the second generation profile HMM. Thus, when the 50 pairs subset was analysed using the iterative `nhmmer v3.1b2` approach a much stricter E-value inclusion threshold was chosen.

2.2.2 Similarity threshold

In order to determine how accurately each homology search algorithm distinguishes true and false homologs a similarity threshold was set (Appendix C.1). All true homologs are expected to be above this threshold. To determine the sensitivity of each alignment algorithm at each PID value a fixed false positive rate (FPR) was used to set the similarity threshold (Eqn. 2.3 A & B). For a FPR of 1 in 10,000 a database with 10,000 sequences would have one 1 false positive (shuffled sequence) above the threshold. The relationship between sensitivity and FPR was tested for a range of 1 in 1250 and 1 in 275,000 FPRs. Total sensitivity was calculated by summing the number of true positives found above the threshold for all PID values and dividing it by the total number of true homologs in the 201 pairs per PID subset. The FPR was chosen to be the lowest value above which very little change in sensitivity was observed.

Sensitivity for each alignment algorithm and PID value was calculated by counting the number of homologs above the threshold (true positives) and dividing by the total number of homologs that were in the database for that gene (true positives and false negatives) (Eqn. 2.3 B). Sensitivity versus PID was graphed for each alignment software and the PID at which sensitivity reached 0.5 is defined as the twilight value for that alignment algorithm.

$$\begin{array}{ll} \textbf{A:} & \text{False Positive Rate} = \text{FPR} = \frac{F_p}{T_n+F_p} \\ \textbf{B:} & \text{Sensitivity} = \frac{T_p}{T_p+F_n} \end{array}$$

T_n is a count of the true negatives (shuffled sequences found below the similarity threshold)

T_p is a count of the true positives (gene sequences found above the similarity threshold)

F_n is a count of the false negatives (gene sequences found below the similarity threshold)

F_p is a count of the false positives (shuffled sequences found above the similarity threshold)

(2.3)

2.2.3 Bootstrapping and 95% confidence interval

In order to gain a measure of variability in the sensitivity a bootstrap analysis was performed. Each pair from the PID subset was labeled as a true positive (above the homology threshold set) or a false negative (below the threshold) for each alignment algorithm. Using an R custom code with built-in bootstrapping function for each PID percent, pairs were selected with replacement (the same number as were available for each PID value) and the sensitivity recalculated (Appendix C.1). This was repeated 10,000 times to create a distribution of values for the sensitivity of each PID percent. The 95% confidence interval (CI) of each distribution was calculated and these bounds used to apply error bars for each PID value and alignment algorithm on the twilight zone sensitivity versus PID graph (*96*, *97*).

Chapter 3

Results

3.1 Genus representative core gene set database

The core gene set database has 10,005,524 ncRNA, 35,524,865 tRNA and 198,628,581 mRNA sequence pairs from 1912 genus representative genomes (Tab. C.3 Link A). In total there were 243,558,970 pairs of sequences. These pairs all had PID values between 0% and 100%. ncRNA sequence pairs (excluding tRNA) had PID values between 21% and 100% whilst tRNA sequence pairs had between 19% and 100% PID (Fig. 3.1). Therefore all ncRNA sequence pairs range between 19-100%. Below 19% there were only mRNA sequence pairs (Fig. 3.1 & A.2). mRNA sequence pairs range in PID from 0% to 100% (Fig. 3.1). Pairs with 0% PID were excluded from the twilight zone analyses. The median PID for each gene type was 72% for tRNA, 62% for ncRNA and 50% for mRNA. The mode for each distribution was very similar (on average +/- 2%) to the median except for ncRNA (excluding tRNA) (Fig. 3.1). The mode for the ncRNA distribution was 71% PID. The shape of the ncRNA distribution shows a secondary smaller mode at 45% PID. The PID distribution for the shuffled variants of each gene has a much lower median and mode (22% and 24% respectively) then both all of the other PID distributions.

There were 142 mRNA, 7 ncRNA and 21 tRNA genes therefore there were just over five times more mRNA genes then ncRNA and tRNA combined. There are 15 mRNA genes with less than 900 sequences from the 1912 genus representative genome set that were removed from the data set to reduce the number mRNA genes. This left 127 mRNA genes. All genes had sequence pairs across multiple PID values. The ratio of ncRNA, tRNA and mRNA in the genus representative core gene subset was unequal for all PID values (Fig. A.2). mRNA genes were 50% or more of the data set for PID values of 68% or less. tRNA was the majority of sequence pairs above this PID value. ncRNA was never the majority for any PID value.

The length of genes in the genus representative core gene subset ranged from 58 to 6306 nucleotides long with a median length of the 783 nucleotides (Fig. 3.2). The length of genes in the database is skewed towards longer lengths.

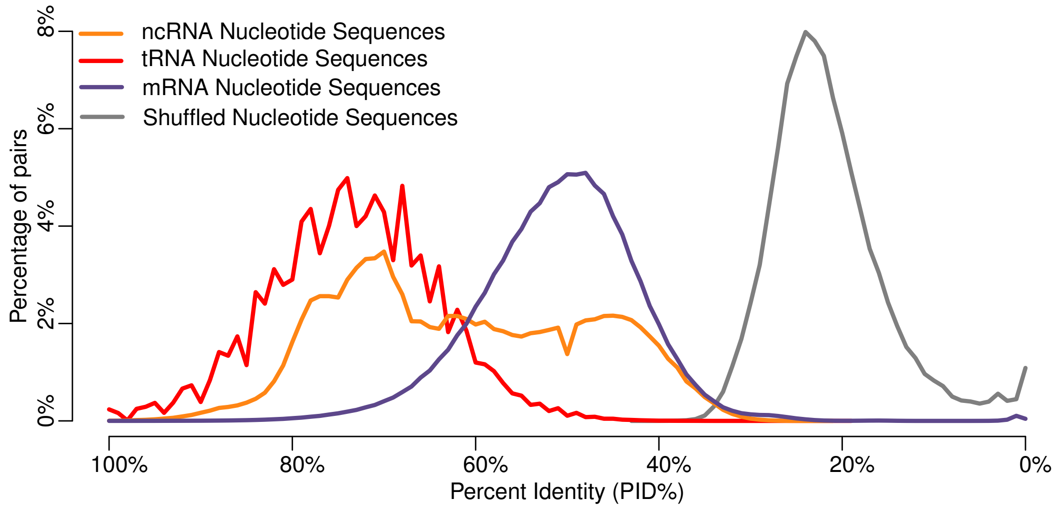


FIGURE 3.1: **PID distributions for each gene type (ncRNA, tRNA, mRNA) and the shuffled version of each gene** The percentage of pairs of the total number of pairs for each group (tRNA, ncRNA, mRNA, amino acid mRNA, shuffled genes) for each PID value. The ncRNA sequence set excludes tRNA sequence pairs.

The median genome GC content for the 1912 genus representative genomes was 55% with a minimum genome GC content of 28% and a maximum of 74% (Fig. 3.2). The G+C content range for all the sequences was 18% to 80%.

The genome set from the NCBI RefSeq database contained a large number of genomes from Proteobacteria, Firmicutes, Bacteriodetes and Actinobacteria (Fig. 3.3). In particular there were 9385 *Escherichia* genomes of which 9324 were *E.coli*. In the genus representative genome set there is one *Escherichia* representative, *E.coli* strain FC10223 (Tab. C.3 Link A). Other commonly used bacterial genera such as *Steptococcus* and *Staphylococcus* also had more then 9,000 genomes. The reduction of genomes to a genus representative genome set reduced these over represented genera down to one genome and this in turn increased the proportion of genomes in the set from phylum where only a small number of genomes have been sequenced (e.g. Caldiserica, Lentisphaerae)(Fig. 3.3)(Tab. C.3 Link D). There was still a large number of Proteobacteria, Firmicutes, Bacteriodetes and Actinobacteria as genomes had been sequenced from multiple different genera for these phylum (Fig. 3.3).

3.1.1 201 pairs per PID subset

The 201 pairs per PID subset had 18,155 sequence pairs from 154 genes from mRNA (127), ncRNA (6) and tRNA (21).

The most abundant gene was ENOG4105BZH (828 pairs, 4.6% of the total number of pairs) and the least abundant gene was SeC (1 pair, <1% of the total number of pairs) (Tab B.4).

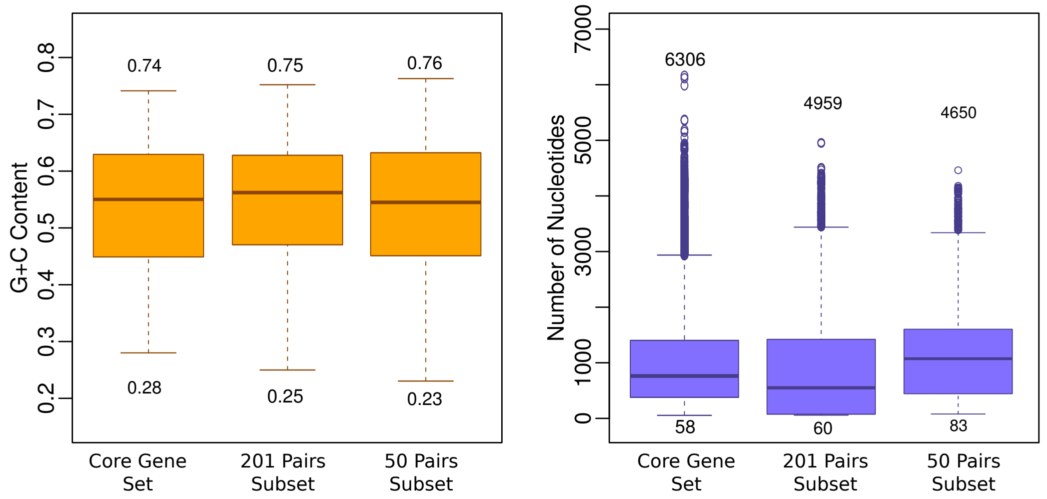


FIGURE 3.2: **G+C content for each genome and length distribution of mRNA and ncRNA sequences in the core gene set, 201 pairs and 50 pairs subset** *Left Panel*: G+C content of each genome calculated using the median G+C content of the core gene sequences for that genome. The G+C content spread is very similar across all three sets. *Right Panel*: Length of each gene in the different sets. The length distributions are also similar although there is more variation in the median, lower quartile and maximum sequence length than the G+C content distributions

Not all PID values in this subset had 201 pairs. PID values of 15% and above had 201 pairs whereas values below 15% had varying numbers of pairs. PID values between 15% and 11% had more then 100 pairs whereas PID values below 11% had 41 pairs or less. There were 168 sequence pairs from 7 mRNA genes for 14% PID, 146 pairs from 7 mRNA genes for 13% PID and 135 pairs from 6 protein genes for 12% PID 3.1. The 11% PID value had 41 pairs from 5 protein genes (Tab. 3.1). 9% and 10% PID had the least number of pairs with 36 and 30 pairs respectivly. All sequence pairs are from the ENOG4105CEY gene except for one pair (ENOG4105DHW) for 9% PID. The 10% PID subset also has the majority of its pairs from the same two genes (14 pairs and 8 pairs respectively) as well as one pair from the ENOG4105BZH gene. From 8% to 1% each PID value had 40 pairs all from the ENOG4105CEY gene.

The proportion of sequence pairs per gene type was unevenly weighted in favour of tRNAs (on average 75%) PID between 100% to 62% (Tab. B.5). Below 62% tRNA's make up 20% or less of the sequence pairs and only populate PID values as low as 32%. From 67% to 1% mRNA sequence pairs are at least half of the sequence pairs for each PID and below 27% there are no ncRNA's. No individual gene has more then 40 pairs per PID value.

The median G+C content per genome, based on the sequences in the 201 pairs per PID subset, is 56% which is 1% greater than the genus representative core gene subset it is derived from (Fig. 3.2). The GC content per genome ranges from 25% to 75% which is similar to the range of the genus representative core gene subset.

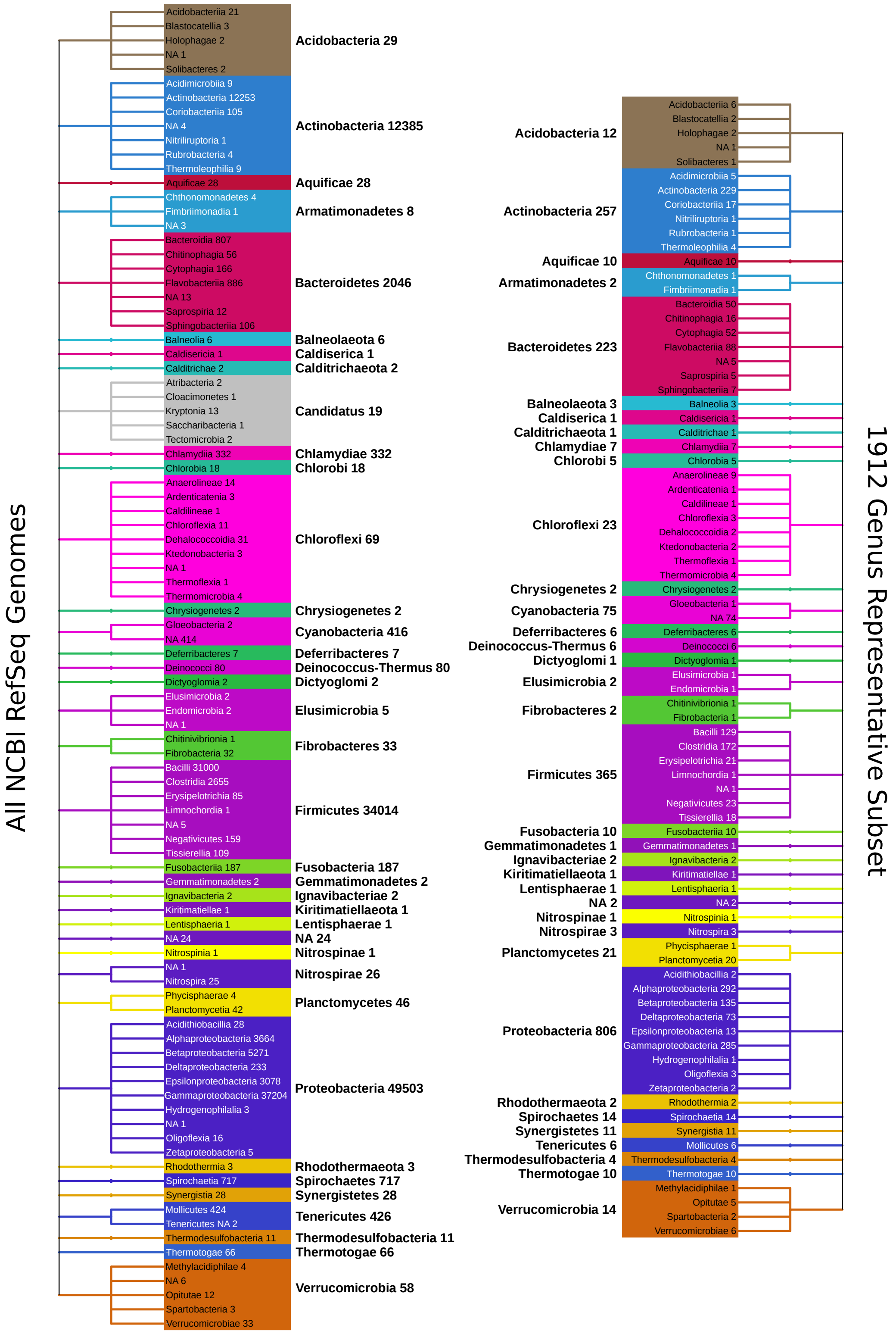


FIGURE 3.3: **Taxonomic phylum and class rank for all NCBI Refseq genomes and the genus representative genome set:** Left Tree: Number of genomes broken down by phylum and class for all genomes in NCBI Refseq
Right Tree: Number of genomes broken down by phylum and class for all genomes in genus representative genome set. A copy of the tree can be found online (Tab. C.3 Link D)

Genes	14%	13%	12%	11%
ENOG4105BZH	✓	✓	✓	
ENOG4105C1H	✓	✓	✓	✓
ENOG4105C90	✓	✓	✓	✓
ENOG4105CEY	✓	✓	✓	✓
ENOG4105CGR		✓		
ENOG4105CWT	✓	✓		
ENOG4105DA0				✓
ENOG4105DHW	✓	✓	✓	✓
ENOG4105KNM	✓		✓	

TABLE 3.1: **Number of protein Genes found for PID values with less then 201 pairs:** All the genes are protein coding genes as the PID of ncRNA does not range below 19%.

The median length for all the sequences in the 201 pairs per PID subset is 552 nucleotides and is 231 nucleotides shorter than the genus representative core gene set (Fig. 3.2). The length of sequences ranges from 60 to 4,959 nucleotides which is a much smaller range than the core gene set. The length of sequences in the 201 pairs per PID subset also has a skew towards longer sequences, with the lower quartile (77) only 17 nucleotides longer than the minimum sequence length. This is likely due to over representation of tRNA genes which are on average 78 nucleotides long. The upper quartile is 1,422 nucleotides long. Therefore approximately three quarters of the data is 1,422 nucleotides in length or less (not exactly three quarters as some longer sequences were separated from the data set and treated as outliers).

3.1.2 50 pairs per PID subset

A 50 pairs per PID dataset with a different ratio of genes was constructed as a further independent small-scale test to validate the 201 pairs subset, due to the tRNA over representation. The subset has 4,578 sequence pairs from 134 different genes from mRNA (127) and ncRNA (7). The tRNA genes were excluded from the subset so a comparison of the twilight zone could be made with the 201 pairs subset twilight zone to determine if the tRNA over representation for high PID values had an effect on the twilight zone.

For the 50 pairs per PID subset the most abundant gene was RF00177 (273 pairs, 6.0% of the total number of pairs) and the least abundant gene was ENOG4105C38 (2 pairs, <1% of the total number of pairs) (Tab B.4).

Not all PID values had 50 pairs of sequences that satisfied the inclusion criteria for all PID values (Methods 2.1.4). PID values of 13% and higher all had 50 sequence pairs. For each PID value below this there was a varying number of pairs; 12% PID had 41 pairs from 5 protein genes, 11% PID had 28 pairs from 5 protein genes and 10% PID had 17 pairs from 3 protein genes (Tab. 3.2). From 9% to 1% PID each had 10 pairs all from ENOG4105CEY.

For 69% to 99% PID ncRNA made up at least 30% of the pairs for each PID value. 100% PID had 8 ncRNA sequence pairs (Fig. 3.13). Between 28% to 68% PID only had a maximum of one fifth of its total pairs as ncRNA genes and below 28% PID there are no ncRNAs (Tab. B.5). No individual gene had more than 10 pairs per PID value.

Genes	12%	11%	10%
ENOG4105C1H	✓	✓	
ENOG4105C90	✓	✓	
ENOG4105CEY	✓	✓	✓
ENOG4105DHW	✓	✓	✓
ENOG4105KNM	✓		
ENOG4105DA0		✓	
ENOG4105BZH			✓

TABLE 3.2: **Number of protein genes found for PID values with less then 50 pairs**

The median G+C content per genome, based on the sequences in the 50 pairs per PID subset, was 54% and was only 1% smaller than the genus representative core gene subset it was derived from (Fig. 3.2). The GC content per genome ranged from 23% to 76% which is similar to the range of the genus representative core gene subset and the 201 pairs per PID subset.

The median length for all the sequences in the subset was 1,119 nucleotides, 336 nucleotides longer than the genus representative core gene subset (Fig. 3.2). The length of sequences ranges from 83 to 4,650 nucleotides which is smaller in range than both the genus representative core gene subset and 201 pairs per PID set. Like the other two subsets the length of sequences in the 50 pairs per PID subset was skewed towards longer lengths (Fig. 3.2).

3.1.3 Duplicates

A small number of ORFs were the top scoring sequence for their genome for two different Eggnog HMMs, meaning they were assigned two different annotations (Tab. B.9). They were above the bit score and length cut offs used to find gene homologs for both HMMs (1/3 maximum bit score possible and between 1/3 and 5/3 the medium length in the alignment). Duplicates were identified after PID selection and removed from the PID subset. All the duplicate annotations were to three HMMs that were different types of GTPases (Tab. B.9). GTPase proteins trigger the GTPase activity of activated G proteins they are bound to in order to halt signalling (*98*).

A visual inspection of the alignment for one sequence (orf178332101) shows that it aligns to other sequences in a highly conserved core region in the middle of the gene that is rich in codons that represent hydrophobic amino acids (Tab. C.1 Link C).

However, it lacks sequence matches to other regions at the start or end of the gene that also show high conservation (also rich in hydrophobic amino acids) (Tab. C.1 Link C) (*99*). Therefore it is only a partial match to the gene but as the sequence is within the bit score and length thresholds applied when determining true homolog core gene sequences it was not excluded. The conserved sections are likely due to all GTPases having similar function and therefore similar functional domains, so a sequence aligning well to the domain in one GTPase will align similarly to another (e.g. the G domain) (*100*). Without further computational analyses it is not possible to determine if these sequences are one of the GTPases or not related to any of the genes.

In the 201 pairs per PID subset there were six different ORFs with duplicate annotations (Tab. B.9). The PID values ranged from 11% to 50% with the majority being below 28%. In the 50 pairs per PID subset there were two different ORFs with duplicate annotations (Tab. B.9). These two ORFs were also found as duplicates in the 201 pairs per PID subset. The PID values ranged from 13% to 28% with the majority being around 20%. All duplicate sequences were removed from all gene sequence sets.

3.1.4 Nucleotide versus protein sequence PID

The protein and nucleotide PID for pairs of mRNA sequences were graphed to determine how they related to each other (Fig. 3.4). The spread of data is monotonic and an almost linear relationship with only a small amount of variation in nucleotide PID (on average around 10%) per amino acid PID value. The association between amino acid and nucleotide PID has a Spearmans correlation of 0.99 (two decimal places)(*101*).

Amino acid and nucleotide PID are surprisingly tightly correlated given the redundancy of the genetic code allowing for different codons to encode the same amino acid. This allows us to see that amino acid PID values in the twilight zone for protein alignments have approximately 20% nucleotide PID or less (Fig. 3.4). Codon preserved shuffled variants of each gene do not have the same PID relationship. They group below and near the intersection between the nucleotide and protein twilight zones from previously published literature (*11*, *29*, *52*–*54*). For higher amino acid PID values the nucleotide PID value tended to be the same or lower than the amino acid PID value (Fig. 3.4). Around 60% amino acid PID this trend changed and the nucleotide PID was the same or higher than the amino acid PID value.

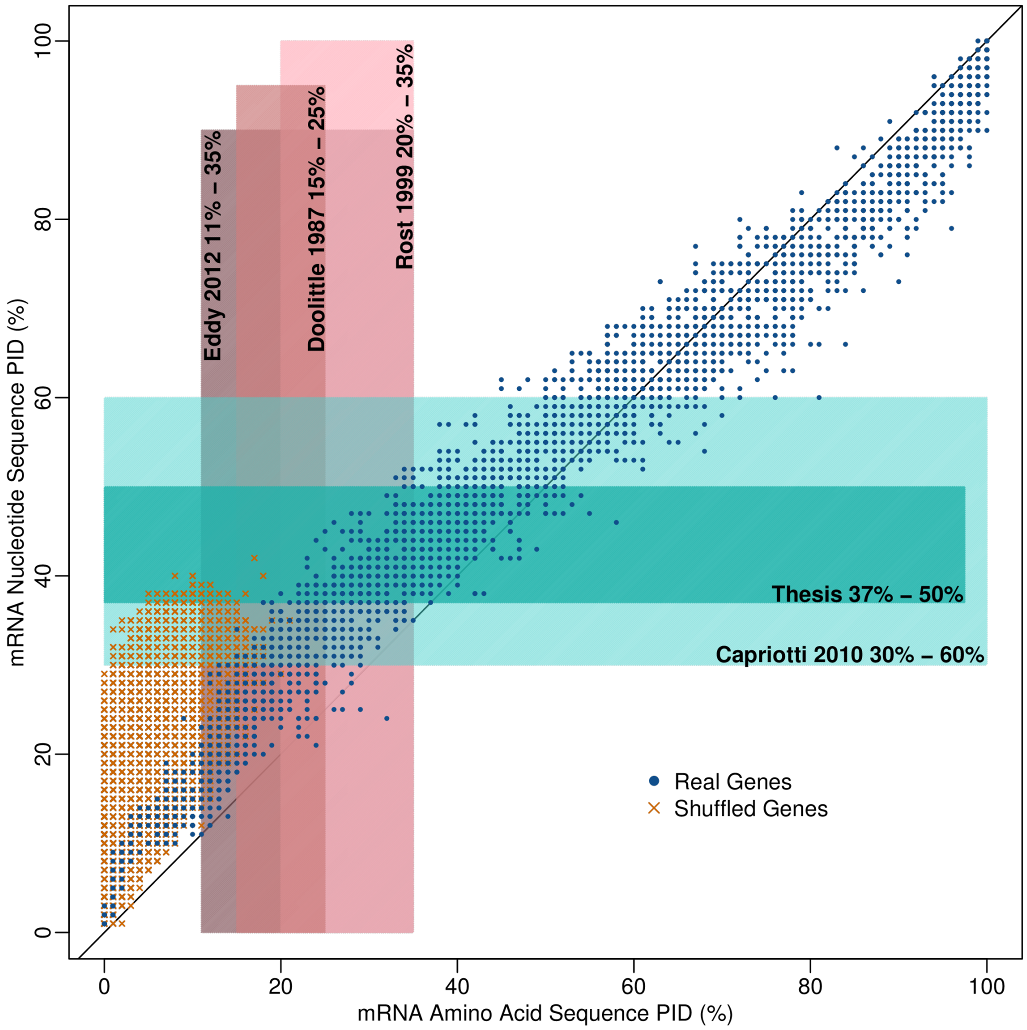


FIGURE 3.4: **PID of protein alignments and the corresponding mRNA alignment for each sequence pair in the 50 pairs subset:** Amino acid PID and nucleotide PID are highly correlated by Spearmans rank correlation coefficient which determines the strength and direction of the association between the two PIDs (0.99 (two decimal places)) (101). All four known twilight zone of homology estimations for protein and nucleotide are shaded on the graph along with the estimation calculated in this thesis using the 50 pairs per PID subset which is labelled as 'Thesis' (29, 52–54).

3.2 Twilight Zone

3.2.1 False positive rate determination

A range of false positive rates (FPR) between 1 false positive out of every 1,000 true negatives above the homology threshold and 1 false positive out of every 200,000 true negatives above the homology threshold were tested to determine how sensitivity correlated to FPR (Fig. 3.5). For both BLASTn and ssearch, sensitivity increased as FPR increased until the FPR reached 0.0001 (1 false positive out of every 10,000 true negatives above the homology threshold) (Fig. 3.5). From a FPR of 0.0001 or larger the sensitivity remained fairly constant. A FPR of 0.0001 was chosen to maximise sensitivity but minimise the number of false positives to include above the homology threshold. The homology threshold for each gene was set to include 28 shuffled sequences (known false positives) for the 201 pairs per PID subset (Eqn. 3.1). The 50 pairs per PID subset homology threshold was set for each gene to include 2 shuffled sequences above it (Eqn. 2.3).

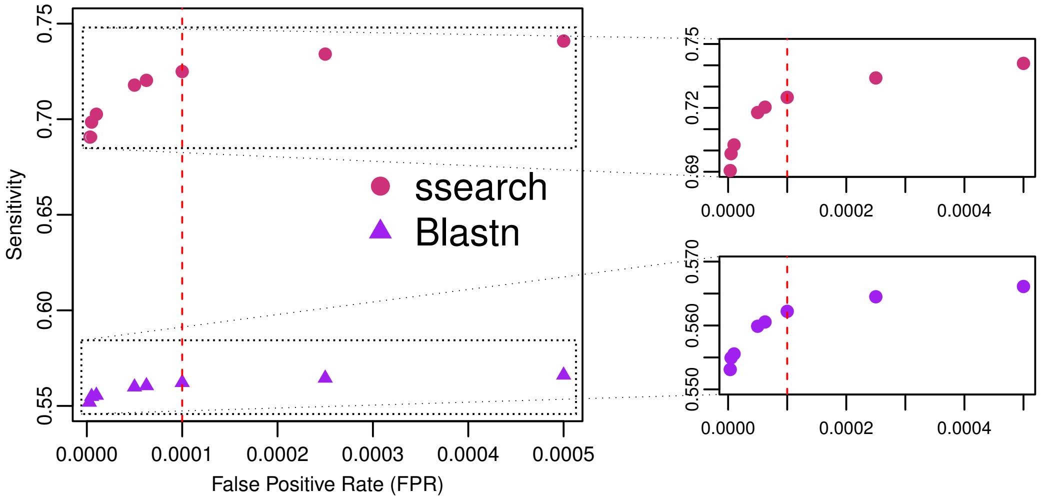


FIGURE 3.5: **The relationship between sensitivity and FPR:** *A)* Depiction of the relationship between sensitivity and FPR for `ssearch` and `BLASTn`. *B)* The relationship between sensitivity and FPR for `ssearch` magnified. *C)* The relationship between sensitivity and FPR for `BLASTn` magnified

$$\begin{aligned} \text{FPR} &= 0.0001 = \frac{F_p}{T_n+F_p} = \frac{F_p}{\text{All shuffled sequences}} \\ 0.0001 \times \text{All shuffled sequences} &= F_p \\ F_p &= 28 \text{ (201 pairs per PID set - 275,351 shuffled sequences = } T_n\text{)} \\ F_p &= 2 \text{ (50 pairs per PID set- 15,996 shuffled sequences = } T_n\text{)} \end{aligned} \tag{3.1}$$

3.2.2 Shuffled sequence database

A database of shuffled versions of the real gene sequences in the core gene subset was generated for both the 201 and 50 pairs per PID subsets. The shuffled database for the 201 pairs per PID subset contained one shuffled copy of every gene in the core dataset (275,351 sequences). This is much larger than the dataset of real homologs which contained 28,982 sequences. The 50 pairs per PID subset shuffled database was much smaller at 15,996 sequences (two shuffled copies of each sequence in the 50 pairs subset). There are 7,998 sequences in the homologous 50 pairs subset.

3.2.3 201 pairs per PID subset

The 201 pairs per PID subset twilight zone was between 40% and 47% for the four core alignment algorithms tested (Fig. 3.6). BLASTn has the highest twilight value of 47% followed by ggsearch which was 44%. ssearch and nhmmer both have the same twilight value of 40%.

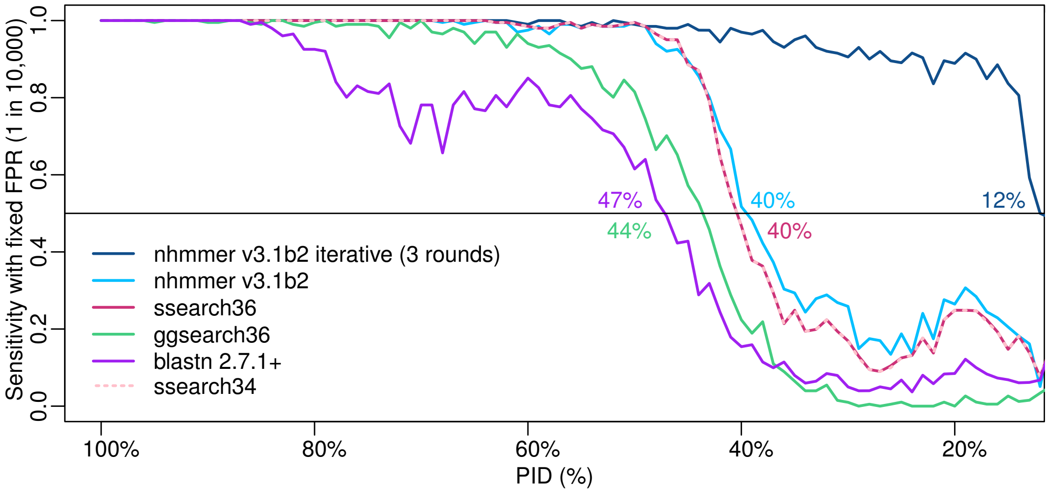


FIGURE 3.6: **Twilight zone of nucleotide homology for the 201 pairs subset:** The twilight value for each alignment software (BLASTn, ggsearch, nhmmer, ssearch36, ssearch34 and nhmmer iterative) was plotted by connecting each sensitivity value for each PID. No trend line or smoothing has been applied.

ssearch34 is a version of ssearch36 that is seven years older however both have the same alignment output and subsequently have the same twilight value (Fig. 3.6).

The nhmmer iterative approach was the best at correctly inferring homology for low PID values (Fig. 3.6). It had a twilight value of 12% which was 28% lower than the next closest twilight value (ssearch and nhmmer, 40%) (Fig. 3.6). Over the twilight zone of the four core alignment algorithms tested (40-47%) the nhmmer iterative approach had on average a sensitivity of 0.97. That is almost double the sensitivity of all the other alignment software for that range.

BLASTn had an unexpected dip in its sensitivity, compared with the other alignment algorithms, between 60% to 80% (Fig. 3.6). This could be potentially due to these PID values having large proportions of tRNA genes. The short length of the tRNA genes increases the chance of finding non-homologous sequences from the database with a similar bit score to true homologs (including their own shuffled variant). The heuristic approach taken by BLASTn means the optimal alignment for a pair of homologs is not always found so the alignment bit score can be lower then it should be (9). As there is a higher chance of finding non-homologous sequences with more similar bit scores to true homologs and the potential for lower bit scores for true homologous sequence pairs, this may mean BLASTn is struggling to separate true tRNA sequences from the shuffled counterparts.

Below 20% PID the four core alignment algorithms show an increase in sensitivity that is unexpected (Fig. A.4).These PID values should be ignored as they likely correspond to protein pairs that have an amino acid PID in the twilight zone so they may not be true homologs (Fig. 3.4).

3.2.4 50 pairs per PID subset

The twilight zone of nucleotide alignment for the 50 pair per PID subset is consistent with the 201 pairs per PID subset and is between 40-49% (Fig. 3.6 & 3.7). BLASTn has the highest twilight value of 49%, followed by ggsearch with 45% (Fig. 3.7). ssearch and nhmmer are both the same and have a twilight value 40% (Fig. 3.7). The twilight value for ssearch34 and ssearch36 are also identical for the 50 pairs per PID subset. The nhmmer iterative approach again has the lowest twilight value (12%). The same increase in sensitivity for PID values of less than 20% is observed in the 50 pairs per PID subset as well (Fig. A.5). Below 39% BLASTn has equal or higher sensitivity values than ggsearch.

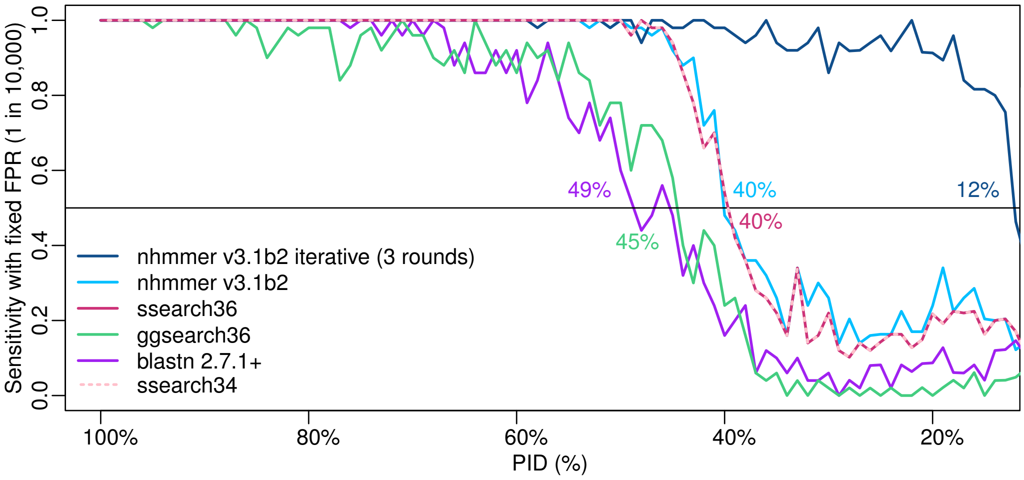


FIGURE 3.7: **Twilight zone of nucleotide homology for the 50 pairs subset:** The twilight value for each alignment software (BLASTn, ggsearch, nhmmer, ssearch36, ssearch34 and nhmmer iterative) was plotted by connecting each sensitivity value for each PID. No trend line or smoothing has been applied.

Bootstrapping gave a measure of accuracy to the sensitivity calculation for each PID value and this allowed identification of a range of PID values for which their error bars included 0.5 sensitivity (aka. the twilight zone). The lowest and highest PID values that include 0.5 sensitivity in their range were used to determine the range of values for the twilight zone of each alignment algorithm individually as well as the overall twilight zone (Fig. 3.8 & 3.13)(Tab. 3.3).

Alignment Algorithm	Bootstrap Twilight Range
`ggsearch`	41-49%
`ssearch`	38-40%
`BLASTn`	43-50%
`nhmmer`	37-40%

TABLE 3.3: **Twilight zone range for each alignment algorithm based on bootstrap error bars**

For all alignment algorithms the error bar size for most PID values is around 0.2 on average until 12% where the error bars in some case almost span the whole range of sensitivities possible (0 to 1) (Fig. 3.8). As shown above, nucleotide PID values of 20% and below correspond to homologous pairs of proteins with around 11% amino acid PID or below (Fig. 3.4). HMMER 3.0 was used (hmmsearch) to find these protein sequences and twilight zone for HMMER 3.0 begins at 11% (Fig. 1.6)(*54*). Therefore mRNA pairs with nucleotide PID of 20% or lower that have been called true homologs in this research may in fact not be due to their amino acid alignment lying in the twilight zone for protein homology. So the sensitivity values for each PID can't be trusted as sequence pairs may be counted as true homologs that are not.

From around 35% to 12% `ggsearch` has very small error bar ranges (<0.1) compared with all other alignment software which shows there is very little variation in the measurement (Fig. 3.8). `nhmmer` iterative on average has the smallest error bars (above 12%).

G+C content

The G+C content of sequence pairs affected the position of the twilight value for each alignment software (Fig. 3.9). Sequence pairs that both have low G+C content (<40%) or high G+C content (>60%) shift the twilight value lower for all four core alignment algorithms (Fig. 3.9). Sequence pairs where one sequence has low G+C content and the other has high G+C content shift the twilight value higher for all four core alignment algorithms.

`ggsearch` appears to be less affected by G+C content than the other three as the change in twilight value is only +/- 1% (Fig. 3.9). The other three alignment algorithms are more affected by G+C content of sequence pairs being the same (G0-G0 or G2-G2) than pairs that have opposite high and low G+C content (G0-G2).

As the 50 pairs per PID value are split between 6 possible G+C content groupings for sequence pairs the number of pairs for each group is very small (Tab. 4.1). This meant a smoothing spline was applied in R with 12 degrees of freedom to identify

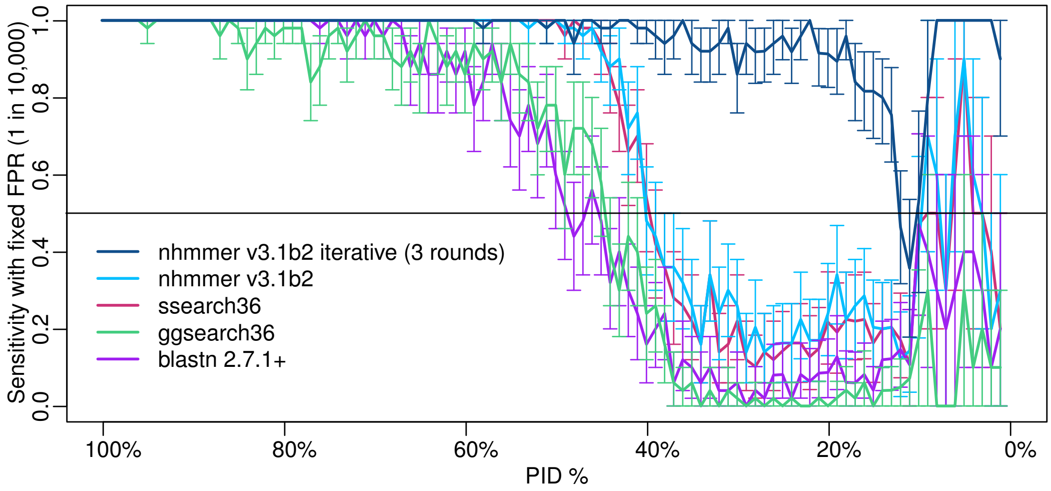


FIGURE 3.8: **Twilight zone of nucleotide homology for the 50 pairs subset bootstrap analyses:** The bootstrap analyses selected sequence pairs with replacement for each PID and calculated the fraction of those chosen that were above the similarity threshold (aka sensitivity). This returned a distribution of sensitivities for each PID and the 95% CI was calculated from the distributions to produce error bars for each alignment algorithm.

the overall trend. The lack of pairs for each PID grouping means the trend lines can be used as an indication of how G+C content affects the twilight zone for each alignment algorithm but conclusions can not be drawn due to it being underpowered.

Length

Length also affects the location of the twilight zone (Fig. 3.10). Sequence pairs that are both short (less than 100 nucleotides) cause the twilight value to shift higher (although the lack of short pairs for each PID means this trend does not hold for the whole PID range for BLASTn and ggsearch) (Fig. 3.10). Sequence pairs between 100 and 1000 nucleotides also cause the twilight value to shift higher (Fig. 3.10). When sequence pairs are 1000 nucleotides or longer the twilight value is shifted lower.

ggsearch is less affected by length than the other three alignment algorithms (Fig. 3.10). The other algorithms are all local alignment tools so this result indicates global alignment algorithms (such as ggsearch) may be less affected by sequence length. More global alignment algorithms would need to be tested to confirm this.

The twilight value of BLASTn is most affected by sequences shorter than 1000 nucleotides long whereas ssearch36 and nhmmer are more affected by sequences that are 1000 nucleotides or longer (Fig. 3.10).

The affect of sequence length can only be considered an indication of the potential impact on the twilight value location. This is because under powered samples may result when splitting 50 pairs between five length groups, and is similar to the issue

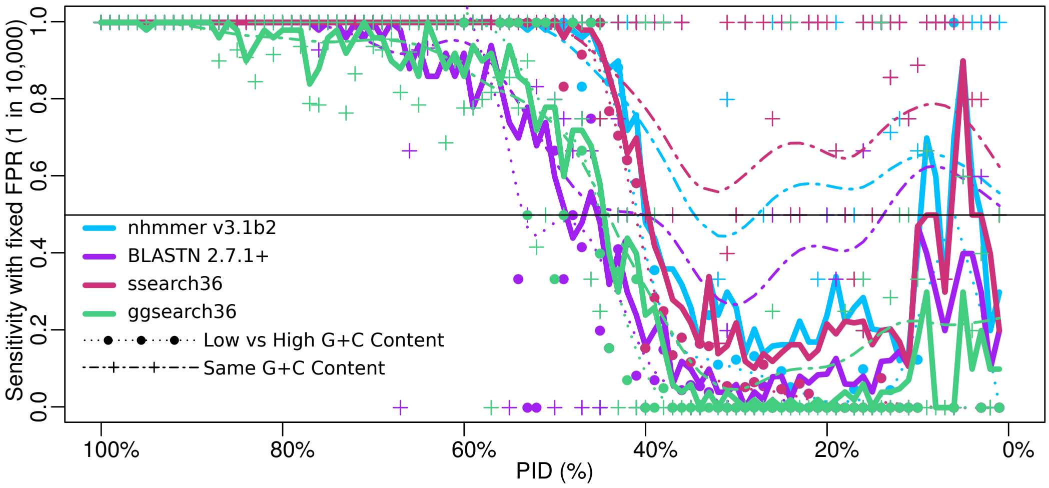


FIGURE 3.9: **Assessing the G+C content effects on the twilight zone for the 50 pairs subset:** Each colour corresponds to an alignment algorithm as denoted in the legend and each line and point type indicates which G+C group the data belongs to. The solid line for each alignment algorithm is the twilight value for all G+C contents combined. A smoothing spline was applied in R with 12 degrees of freedom to each G+C content group to identify the overall trend. This was done to smooth out the effect of having a small number of pairs for each group.
Key: Large filled circles with a dotted line indicate sequence pairs with G0-G2 G+C content, Plus shaped points with dash-dot perforated lines indicate sequence pairs with G0-G0 and G2-G2 G+C content. The color denotes the corresponding alignment software tool.

presented by G+C content (Tab. 4.1). There are only five length groups, not six, because there can be no short versus long sequence pairs (L0-L2) due to the length restrictions placed when finding the core gene set.

3.2.5 Computational timings

For the 50 pairs per PID subset all alignments were done in parallel (each group of sequences for one gene is independent to the alignment of another gene so can be run in parallel) to decrease the time taken to align all the sequences to the required algorithms (Tab. C.1). For the 50 pairs subset all three alignment procedures (four core alignment algorithms, `nhmmer` iterative, `ssearch34`) were 'clock' timed using /usr/bin/time (Tab. 3.4). The four core alignment algorithms took the least amount of time and the smallest number of parallel threads were utilized however each alignment algorithm used the most CPU (computational cores) of all the alignment algorithms (Tab. 3.4). `ssearch34` was the slowest however it can only run on one CPU whereas all the other software can run on any specified number of CPUs (Tab. 3.4).

Not all computational runs were timed for the 200 pairs per PID subset so it is not possible to make direct comparisons for every alignment software. `nhmmer` iterative

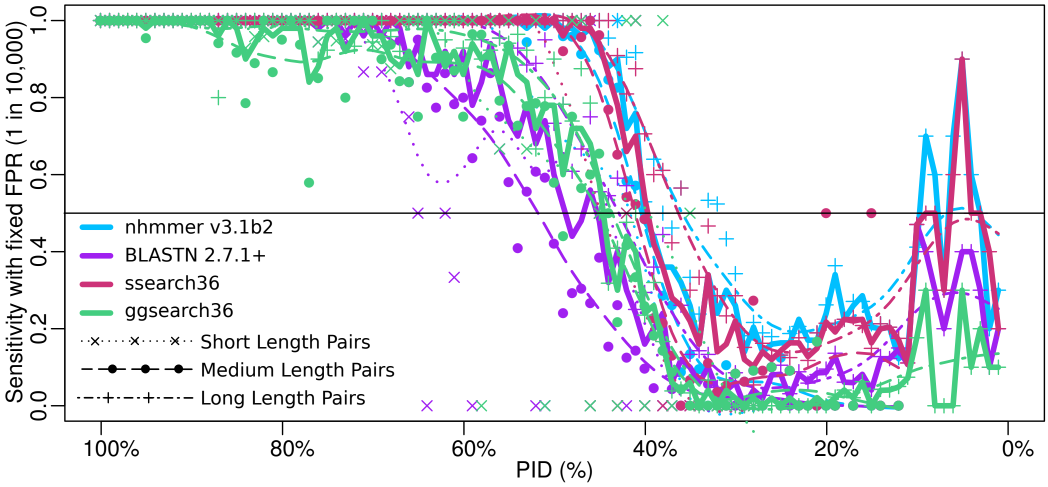


FIGURE 3.10: **Assessing the length effects on the twilight zone for the 50 pairs subset:** Each colour corresponds to an alignment algorithm as denoted in the legend and each line and point type indicates which sequence length group the data belongs to. The solid line for each alignment algorithm is the twilight value for all sequence lengths combined. A smoothing spline was applied in R with 12 degrees of freedom to each length group to identify the overall trend. This was done to smooth out the effect of having a small number of pairs for each group.
Key: Cross shaped points with dotted lines indicate short sequence pairs, Plus shaped points and dash-dot perforated lines indicate long sequence pairs, Large filled circles with a dashed line indicate medium length sequence pairs. The color denotes the corresponding alignment software tool.

took 168 hours using 15 cores at 3 CPU for the 201 pairs per PID subset. ssearch34 was run in two different ways for the 201 pairs per PID subset due to the amount of time it takes to run. Initially it ran for 679 hours on 1 core at 1 CPU. Then the remaining 118 models were run in parallel on 8 cores at 1 CPU for 908 hours. In total ssearch34 took 1,587 hours (66 Days) for the 201 pairs subset which is a long time compared with current day versions.

Alignment Software	CPU	Clock Time
ssearch36, ggsearch, BLASTn, nhmmer	15	5 Hours
nhmmer iterative (3 rounds)	30	12 Hours
ssearch34 (old version of ssearch36)	20	16 Hours

TABLE 3.4: **Computational timings of different alignment software:** The reported time each process took to run is "wall time" not CPU time. e.g. The four core alignment algorithms tested (nhmmer, ssearch, BLASTn and ggsearch) took 5 hours in total to run all four algorithms for the 50 pairs subset using five parallel runs of the code each using 3 CPU therefore 15 CPU total.

3.2.6 PID distributions of shuffled or random Sequences

Shuffled gene control set

The shuffled gene set sequences were split into G+C content and sequence length groups to determine what effect they had on the PID distribution. The modes of each distribution range between 19-29% PID and the maximum value for each distribution ranges between 33-43% PID (Fig. A.6 A-I).

The largest maximum PID value for the G+C content groups was 43% for sequence pairs where both sequence have high or low G+C content (Fig. 3.11). The G+C content group with the lowest maximum PID value was sequence pairs where one sequence has low and the other has high G+C content (33%) (Fig. 3.11). The modes of the two distributions also have a similar difference (22% to 29%). This shows that G+C content can affect the background PID values expected. It also shows that PID values of 43% or lower are possible for randomly aligned sequences.

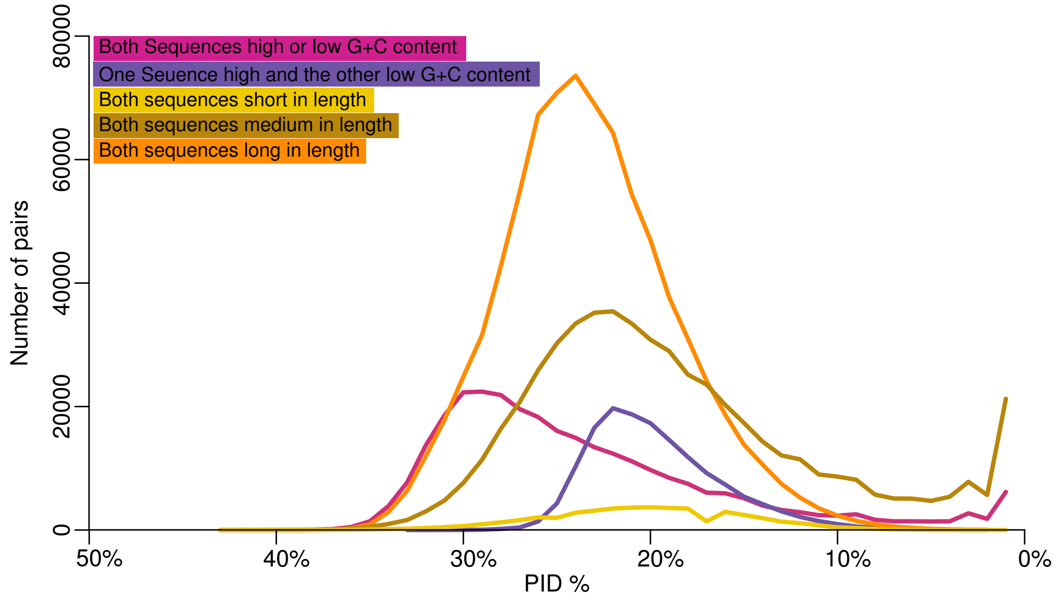


FIGURE 3.11: **PID distributions of shuffled gene sequences based on G+C content and sequence length:** The distribution for short sequences (L0-L0, yellow) has a mode of 20% and a maximum value of 43%. The distribution of medium length sequences (L1-L1, brown) has a mode of 22% and a maximum value of 42%. The distribution of long length sequences (L2-L2, orange) has a mode of 24% and a maximum value of 39%. The distribution of sequence with the same high or low G+C content (G0-G0 or G2-G2, pink) has a mode of 29% and a maximum value of 43%. The distribution of sequences with one high and one low G+C content sequence (G0-G2, purple) has a mode of 22% and a maximum value of 33%.

Key: L0 is sequences less then 100 nucleotides long. L1 is sequences between 100 and 999 nucleotides long. L2 is sequences 1000 nucleotides or longer. G0 is sequences with 40% or less G+C content. G2 is sequences of 60% or more G+C content.

The length of sequence pairs also causes variation in the range of PID values for each sequence length distribution. Short sequence length pairs had the highest maximum value (43%) however long sequence pairs had the highest mode (24%) (Fig. 3.11). The maximum value for long sequence pairs is 39%. The number of pairs for each sequence groups is not even, as highlighted by the distributions, with the long sequence group having more sequences than any other G+C content or sequence length group (Fig. A.6). There is a smaller range of modes (20-24%, where 20% is the mode for short sequence pairs) and maximum values (38-43%) for sequence length groups than G+C content groups (Fig. 3.11). This indicated G+C content had a larger effect on PID than sequence length.

Random sequence set

The random sequence set had sequence lengths chosen to mimic distribution of sequence lengths in the core gene set and had two different G+C contents; 28% or 74%. These two G+C contents were chosen as they are the highest and lowest genome G+C content for the genus representative set based on the core genes (Fig. 3.2).

The sequences were split into the same length groups as the shuffled sequence set. The four distributions show that G+C content has a larger effect on PID then sequence length (Fig. 3.12). The difference in the modes between short (less then 1000 nucleotides long) sequences that had either the same or opposite G+C content is 17% (40-57%) (Fig. 3.12). The difference in modes is similar for long sequences (16%, 41-57%). There is a 1% difference in the modes between long sequences with opposite G+C content and short sequences with opposite G+C content (40-41%). There is no difference in the modes between sequences of long and short length of the same G+C content (both 57%). The maximum value for each PID distribution also follows the same pattern as the mode, where very little difference is seen due to sequence length. The highest PID value for all distributions was 89% for a sequence pair that was long and had the same G+C content (Fig. 3.12). This shows that for this sequence set, length does not affect PID whereas G+C content does.

It should be noted that, whilst the equation used to calculate PID was the same as for the shuffled gene and PID subsets, the method in which alignments were made is not the same. As the randomly generated sequence set have no corresponding HMM or CM ggsearch was used to align these instead, however the shuffled sequences do have corresponding HMMs or CMs, therefore the hmmalign or cmalign were used for these. The PID values are likely to be higher for this set of sequences due to the method (Eqn. 2.2). Thus, a direct comparison of PID values cannot be made with the PID subsets and shuffled gene sequences. Therefore the PID distributions for the shuffled gene control set is the only null distribution that can be directly compared to understand underlying chance sequence similarity.

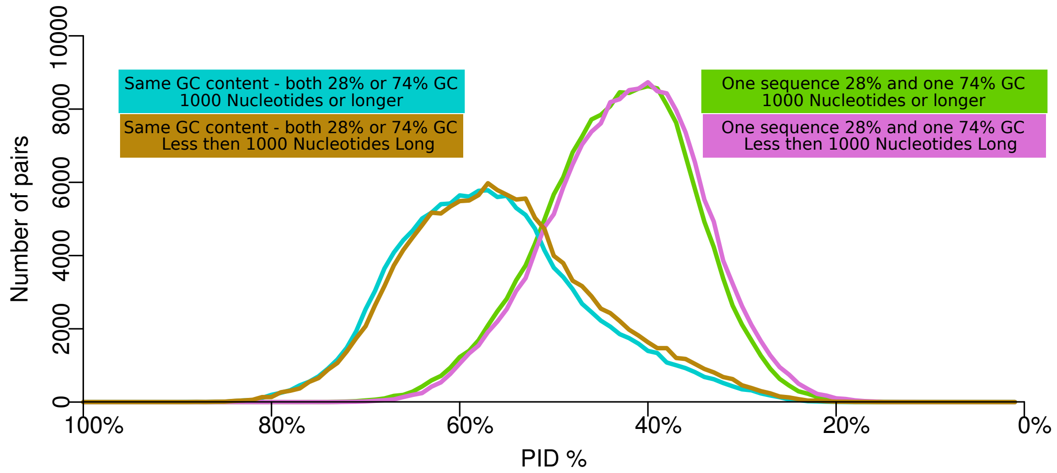


FIGURE 3.12: **PID distribution of randomly generated nucleotide sequences:** The four distributions represent different length and G+C combinations and their effect on the PID values possible for random sequences (background noise to expect)

3.2.7 Final twilight zone estimation

The overall twilight zone for the four core alignment algorithms is between 37-50% PID based on the data from the 50 pairs subset (Fig. 3.13). The bootstrap error bars were used to identify the range of PIDs in the twilight zone (sensitivity of 0.5) for each alignment software (Tab. 3.3). `nhmmer` (37-40%) and `ssearch` (38-40%) have fairly small twilight zone ranges whereas `ggsearch` (41-49%) and `BLASTn` (43-50%) have much larger variation in the twilight zone (Fig. 3.13). The ranges for `nhmmer` and `ssearch` overlap as do the ranges of `BLASTn` and `ggsearch` (Fig. 3.13). There is no overlap between these two groups of alignment algorithms.

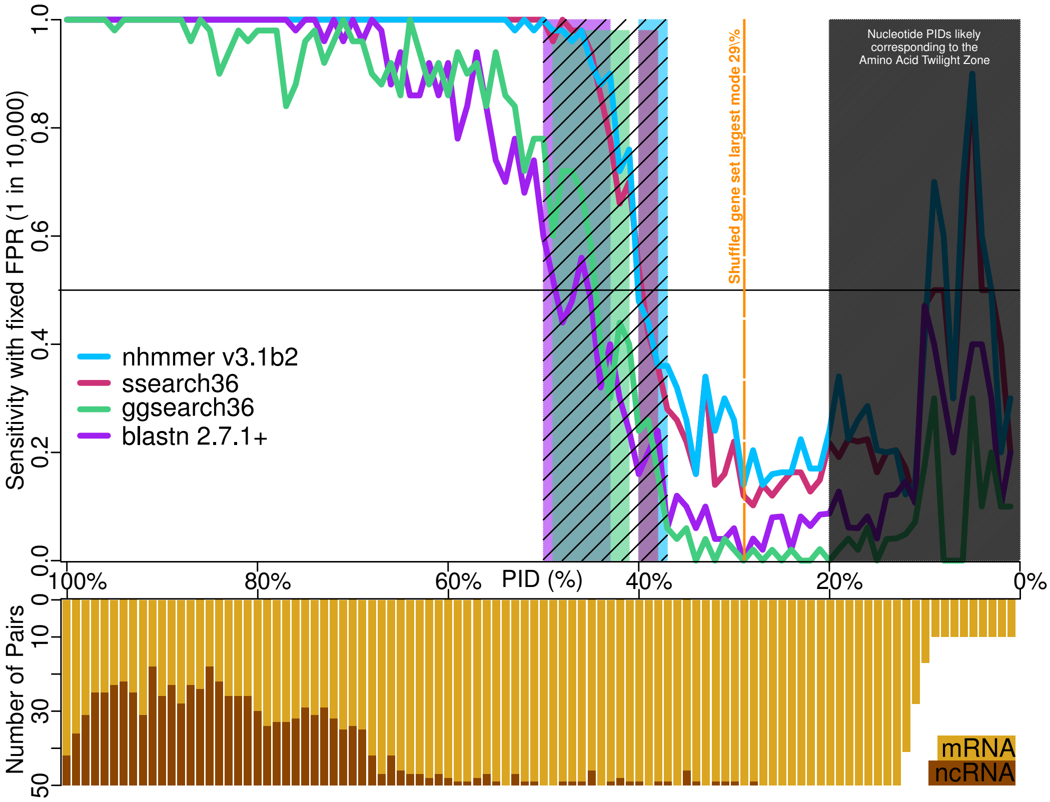


FIGURE 3.13: **Twilight zone for nucleotide homology summary:** Graph summarises all the data used to determine the twilight zone range for each alignment algorithm (*top graph*) as well as the proportion of ncRNA and mRNA sequence pairs for each PID (*bottom graph*). The largest mode for a distribution of PID from the shuffled sequence set (orange dashed line) shows the PID value around which most random sequences are found (Fig. 3.11). It corresponds to the mode for shuffled sequences pairs with G0-G0 or G2-G2 G+C content. The black shaded region identifies the nucleotide PID range in which protein sequence pairs are likely in the amino acid homology twilight zone. Therefore these PID values were excluded.

Chapter 4

Discussion

4.1 Twilight Zone

This research set out to better define the twilight zone for nucleotide homology. The twilight zone is defined as the range of sequence similarity for which the relationship between similarity and homology breaks down to the extent it is not possible to accurately infer homology using sequence similarity (*29*, *52*). A better definition of the nucleotide homology twilight zone will allow for further understanding of the limitations of homology search algorithms and inform researchers when further evidence, other then a significant alignment score, is required to infer homology between two sequences.

4.1.1 50 pairs subset

The twilight zone for nucleotide homology was found to be between 37-50% for the 50 pairs per PID subset (Fig. 3.13. This agrees within the zone calculated by Capriotti *et al.* 2010 which was 30-60%, although the zone found by this research is smaller (*53*). The twilight zone calculated is lower than expected based on discussion with experts in homology search algorithms, although the order of each algorithms twilight zone is as expected (Personal Correspondence with S.Eddy & T.Wheeler) (*54*).

The twilight zones for `ssearch36` (38-40%) and `nhmmer` (37-40%) were almost indistinguishable and had the lowest twilight zones of the four single-sequence query algorithms (Fig. 3.13)(Tab. 3.3). Their similarity is due to the power of HMMs used in `nhmmer` being reduced due to the profile being generated from one query sequence. The query HMM therefore lacked information about the probabilities of a match, mismatch or gap at each position so is similar to aligning with a query sequence rather then a probabilistic model(*10*, *31*). `BLASTn` had the highest twilight zone (43-50%) as expected based on previous research (Fig. 3.13)(Tab. 3.3) (*52*). This is because `BLASTn` uses k-mer (word) based heuristics to accelerate homology searches of large databases so the optimal alignment for a pair of true homologs may not be found (*41*, 42). This leads to bit scores of true homologs being more similar to non-homolog bit scores with high similarity by chance (34).

ggsearch was expected to have a lower twilight zone than ssearch36 which was confirmed by this research (41-49%) (Fig. 3.13)(Tab. 3.3). This is because ggsearch uses global alignment whereas ssearch36 aligns locally conserved regions (4, 5, 39). So the score for ggsearch is likely to be less than or equal to the score from ssearch36 (for the same scoring scheme) due to penalising dissimilar regions that are aligned by ssearch.

Between 65-85% ggsearch had a lower sensitivity then BLASTn although there is no obvious reason for this (Fig. 3.7). The same range for the 201 pairs subset shows ggsearch to be more sensitive than BLASTn for all PID values (Fig. 3.6). The error bars for sensitivity at each PID for the 50 pairs subset are much larger for ggsearch in this range than the other alignment algorithms (Fig. 3.8). In most cases the error bars reach a sensitivity of 1 (the sensitivity values of all the other alignment algorithms over most of this range) or overlap with the BLASTn error bars. As the sensitivity error bars overlap this means the difference between BLASTn and ggsearch in this range is not significant and is within the range of variation for the sensitivity calculation for that PID value.

The number of pairs for each PID could explain this small sensitivity difference between ggsearch and the rest of the four core alignment algorithms. The 50 pairs subset has 50 different potential sensitivity values for each PID and the 201 pairs subset has 201 potential sensitivity values. Therefore, missing 5 homologs in the 50 pairs subset causes a more drastic change in sensitivity (0.9) than the 201 pairs subset (0.98), as 10% of the pairs are below the similarity threshold (Fig. 4.1). This results in the 50 pairs subset placing more weight on a missing homologous pair for a PID value then the 201 pairs subset and has less statistical power than the 201 pairs subset. The 50 pairs subset was kept small to decrease the compute time of the analyses after an error in the proportion of genes for each PID was found in the 201 pairs subset (Tab. B.5).

4.1.2 201 pairs subset

The 201 pairs subset has an unbalanced proportion of tRNAs for PID values 62% and above (Tab. B.5). However, the twilight value for each alignment algorithm is within the range found for the 50 pairs subset that has no tRNA pairs in its subset (ssearch 40%, nhmmer 40%, ggsearch 44% and BLASTn 47%) (Fig. 3.6 & 3.13).

tRNAs presented significant issues for the structural alignment algorithms used in the work of Gardner *et al.* 2007. This BRaliBase dent problem was avoided in this study as the alignment algorithms do not use structure to align sequence pairs, just

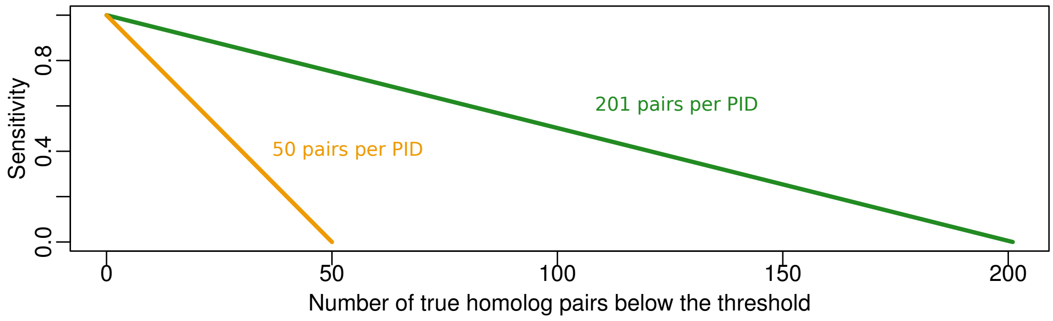


FIGURE 4.1: **Illustration of the difference in sensitivity for the 50 and 201 pairs subset:** The sensitivity of the 50 pairs subset decays faster and is always smaller then the 201 pairs subset for the same number of true homolog pairs below the similarity threshold. This shows that the 50 pairs subset is more sensitive to a missing homolog pair then the 201 pairs subset.

sequence similarity (*64*). However, having tRNA in abundance does cause a different problem due to their short length (78 nucleotides on average). Short sequences have an increased chance of non-homologous sequences aligning with a similar bit score to true homologs than longer sequences (*20*, *29*). As shown for the 50 pairs subset this causes the twilight value to shift higher (by up to 3%) for each alignment software (except ggsearch (Fig. 3.10).

BLASTn has a dip in sensitivity between 60-80%, similar to the BRaliBase dent, that isn't seen in any of the other alignment algorithms for the 201 pairs subset (Fig. 3.6)(*64*). This may be due to over representation of tRNA in this range combined with the relatively short length of tRNA sequences compared with most the other genes (3/4 of 201 pairs subset sequences have length of 77 nucleotides or longer) (Fig. 3.2)(Tab. B.10). The short length of tRNAs, and the fact that the shuffled database contains a large number of shuffled tRNA sequences (37,768) may mean that the BLASTn heuristic approach struggles to separate the di-nucleotide shuffled tRNA sequences from the true tRNA homologs. As the tRNAs make up the majority of the pairs in this region, their length biases the combined sensitivity outcome for each PID value. This would explain why the dip in sensitivity is nonexistent for the 50 pairs subset that has no tRNAs. Whilst there is no caveat on length of pairs selected for each PID value, by restricting each gene to be no more than 20% of the pairs, the chance of diversity in the length of pairs that are selected is increased.

4.1.3 Old version of ssearch

ssearch34 and ssearch36 have the exact same sensitivity for each PID for both the 50 and 201 pairs subset (Fig. 3.6 & 3.6). This is because both use a dynamic implementation of Smith-Waterman to determine the optimum alignment and thus the underlying mechanisms for scoring the alignments is the same (*30*, *39*). ssearch36 is a lot faster due to the implementation of Streaming SIMD Extensions 2 (SSE2), a fast mechanism for calculating score matrices in parallel and determining the possible local alignments for each pairing (Tab. 3.4)(*40*). `ssearch36` also allows for multiple CPU threads to be used by the algorithm. A direct comparison of the 'clock' time each version of `ssearch` took to run cannot be made due to utilising a different number of parallel threads and CPU (Tab. 3.4). However `ssearch34` took three months to run the 201 pairs subset compared with all four core alignment algorithms that took 7.5 days. Regardless of the comparability of the times, that is a much shorter time frame.

In hindsight, `ssearch` was not the best algorithm to use for a comparison to the old version due to the underlying mathematics not changing. Instead, the same algorithm is employed, but it is done in parallel (*40*). Older versions of `BLASTn` are likely to have different results as the underlying mathematics have changed over the years which will give different twilight zones (*9*, *38*). Future work could compare older versions of `BLASTn`, that are compatible with a Linux Ubuntu 17.10 environment, to identify if new versions are more accurate at detecting homologs in the twilight zone then older versions.

4.1.4 nhmmer iterative approach

`nhmmer` was utilised in an iterative approach where multiple sequences from the database that had a significant E-value were included alongside the query sequence to build the HMM to be used for the alignment. Two database search rounds, the first with a single query HMM and the second with a HMM built from the sequence alignment of significant sequence matches from the first round, were used to build the final HMM. This HMM was then used to search the database for sequence matches.

The `nhmmer` iterative approach has a much lower twilight zone than any of the other alignment algorithms for both the 201 and 50 pairs subset (12% for both) (Fig. 3.6 & 3.7). The low twilight zone is not unexpected due to the power of multiple sequence alignment based profile HMMs (*61*). The probability based scoring scheme for `nhmmer` iterative is more powerful than the +4 match, -5 mismatch scoring scheme used by `BLASTn`, `ssearch36` and `ggsearch` as it can penalise matches and mismatches at residues in the sequence differently depending on how well they are conserved, or not, amongst the sequences used to build the hmm (*10*, *31*, *43*).

Whilst the iterative `nhmmer` approach appears superior, there is a problem with the calculation of its twilight zone. The twilight zone for the `nhmmer` iterative approach lies in the nucleotide PID range that corresponds to the protein homology twilight zone determined by Sean Eddy for HMMER 3.0 (Fig. 3.4) (*54*). Tools from the HMMER 3.0 suite were used to determine the set of homologous sequences for each protein gene in the core gene set (*46*). Therefore pairs of sequences that were determined to be homologous that are found in this twilight zone may in fact not be. These pairs should have been excluded or structurally verified to determine if they were true homologs. As only sequence similarity has been used to infer homology, there is a lower confidence in the sequence pairs in this PID range being true homologs so the measure of sensitivity for these PID values is possibly erroneous. Further work to ensure pairs of sequences with very low PID are in fact true homologs is required (e.g. determining structural similarity, genomic context, shared motifs or domains, and functional assays).

4.1.5 Confidence in the FPR

In order to calculate a measure of sensitivity for each PID value, the false positive rate (FPR) was set to 1 in 10,000 (Fig. 3.5)(Eqn. 3.1). This allowed a bit score threshold to be set for each gene in order to determine which sequence pairs would be considered homologous based on the sequence similarity for that alignment algorithm (Eqn. 2.3). The distribution of false positive bit scores is dependent on the sequences in the database and the size. If there are more sequences in the database, high bit scores would occur at a greater frequency by chance based on the E-value (Eqn. 1.1)(*20*). Most databases are very large (e.g. BLAST had 51,150,124 nucleotide sequences in its database on the 23rd of March 2019). Therefore the number of false positives with high bit scores by chance would be greater than the number found for the databases in this body of research (275,351 for the 201 pairs analyses and 15,996 for the 50 pairs analyses). The database sizes were chosen to minimise the amount of compute time for the alignment algorithms. This may mean the number of false positives found above the score threshold are being underestimated when compared to larger databases that are generally used. As FPR influences sensitivity a larger database may give more confidence in the bit score threshold chosen for each gene. Future research will determine the balance point between increasing database size and compute time for alignment algorithms by running test sets of various sizes and calculating how time scales with increasing database size. This will allow future research to utilise the largest database the time available for analyses.

4.1.6 PID calculation

The PID values were determined for each sequence pair in the data set by comparing their similarity to reference alignments. This reference alignment was made by structurally aligning ncRNAs with Rfam CMs and reverse translating protein alignments to EggNOG HMMs. This means the sequences in the alignment are aligned using information in addition to the nucleotide sequence. This was done to obtain an alignment that reflects a reliable estimate of the potential evolutionary events that generated the sequence divergence rather than just being influenced by sequence composition that can have biases (e.g. G+C content). Therefore the PID for each pair of sequences is a representation of their evolutionary divergence, not just sequence composition similarity (Fig. A.3) (*20*). This, however means the PID may be a lower estimation of the sequence similarity in some cases compared with aligning two sequences with respect to each other and maximising similarity.

To examine this further a small sample of sequence pairs, two mRNA and two ncRNA, were chosen from a range of PID values (40-80%). They were aligned using `ggsearch` with the same parameters as for the twilight zone alignment to see how much the PID differed (Tab. B.10 & C.2). The biggest increase was of 17% PID and the smallest was of 3% PID. This shows that calculating the PIDs using direct pairwise comparison, rather than the method used for this thesis, would likely cause the twilight zone to shift higher for each alignment algorithm. mRNA has a further difference in how nucleotide PID is calculated as codons are aligned in-frame of the amino acid alignment. It was calculated this way as natural selection generally acts on the amino acid sequence for proteins not the nucleotide sequence. Therefore, in order for the nucleotide alignment to reflect evolutionary events the nucleotide alignment should reflect that of the protein sequence. This can give a lower nucleotide PID value for each mRNA as it disregards the actual sequence of nucleotides (Tab. B.10). Therefore, the twilight zone range found in this thesis is likely influenced by the way PID is determined for each pair. This may explain why the twilight zone is lower then experts expected (Personal Correspondence with S.Eddy & T.Wheeler) (*54*).

The strong correlation between nucleotide and amino acid PID

There was an unexpected strong correlation between nucleotide and amino acid PID (Spearmans ranked correlation coefficient of 0.99)(Fig. 3.4). It was surprising due to the redundancy in the genetic code allowing different codons to encode the same amino acid, hence allowing the nucleotide sequence to vary more then the amino acid sequence. Some amino acids have one codon and therefore will have 100% PID for each codon in the sequence alignment where these matching amino acids are found. Other amino acids have more codons so the PID of two identical amino acid sequences can in theory range from 100% to 0% nucleotide PID depending on the amino acids in the sequence. It is unlikely to be 0% as that is only possible if the sequence only contains serine. However this full range is not seen for the 50 pairs dataset as the nucleotide PID range is 90-100% for identical amino acid sequences (100% amino acid PID) (Fig. 3.4). This is the same for every amino acid PID value where the range of nucleotide PIDs is in general approximately 10% (Fig. 3.4). For higher amino acid PID values the nucleotide PID tends to be the same or lower than the amino acid PID value (Fig. 3.4). At around 60% this trend changes and the nucleotide PID is the same or higher than the amino acid PID value. This is because mismatched codons for the same amino acid can have nucleotide similarities

≤66.7%. Therefore, while the amino acids are different, a percentage of similarity can occur for mismatched amino acids at the codon level (24% on average). This penalises the amount of change between mismatched amino acids at the nucleotide level differently depending on if one, two or three mutations have occurred to cause the amino acid change. This means the nucleotide sequence gives more information on the amount of divergence between homologs then the amino acid sequence.

Accessing the influence of G+C content and sequence length

The shuffled and randomly generated sequence sets were created as a set of control sequences to show the distribution of PIDs expected for aligned unrelated sequences as well as to help identify the effect of sequence composition on sequence similarity (Fig. 3.11, 3.12 & 4.2). Of particular interest was the effect of sequence composition factors G+C content and sequence length on the PID distribution.

The shuffled sequences were generated by di-nucleotide or codon shuffling the 50 pairs subset sequences. These sequences were then aligned to the CM or HMM for the gene they originated from. The CMs for two genes, 16s rRNA and 23s rRNA, would not align the shuffled sequences to the CM. This occurs when the sequences lack complexity or the sequence cannot be classified by the CM. So these two genes were excluded from the shuffled set. The random sequences were generated by random arrangement of sequences at specified length with either 24% or 78% G+C content. They were then grouped by length (short, medium, long) and aligned in pairs using `ggsearch`. Therefore, the shuffled sequences PID was calculated from a reference alignment of all shuffled sequences for a gene and the random sequences from individual pairwise alignments that were maximised for sequence similarity between the sequences (Fig. 4.2). This means the PIDs for the shuffled sequences will be lower than the random sequences.

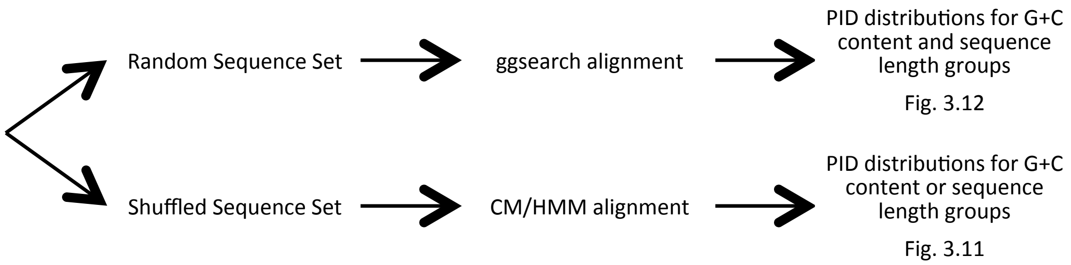

Random Sequence Set
ggsearch alignment
PID distributions for G+C content and sequence length groups
Fig. 3.12
Shuffled Sequence Set
CM/HMM alignment
PID distributions for G+C content or sequence length groups
Fig. 3.11

FIGURE 4.2: **Overview of shuffled and random sequence set analyses:** Highlighting the different alignment processes for each set and the resulting distributions. As the randomly generated set had more sequence pairs than the shuffled set, pairs could be separated by both G+C content and sequence length. The shuffled sequences could not.

The highest PID of a shuffled sequence pair is 43% whereas the highest random sequence pair is 89% (Fig. 3.11 & 3.12). There is also a large difference between the modes (lowest random sequence mode is 41% and the highest shuffled sequence mode is 29%) (Fig. 3.11 & 3.12). The difference in the PID values is likely due to how they were aligned to calculate PID and the low complexity features (a lot of mono-nucleotide segments) in the random sequences increasing sequence similarity compared with the shuffled sequences (Tab. C.3 Link F). In future research a more stringent approach to random sequence generation where sequence complexity is considered will be utilised (*102*).

The shuffled and random sequences distributions also highlights how the method for PID calculation used in this research likely causes a lower PID estimate when compared with pairwise aligning sequences (Fig. 3.11 & 3.12). As the sequences used for each set are not the same, definitive conclusions can not be drawn as to how much of the difference is due to the alignment procedure. Comparison of the current PIDs of the shuffled sequences to the PIDs generated using the random sequences alignment method would be required to make an actual comparison. A detailed check of one pair of shuffled mRNA sequences (two different shuffled variations of the same gene so the sequence length and G+C content is the same) showed the nucleotide PID increased by 9%. The increase in the nucleotide PID is because the nucleotide sequences have the flexibility to align outside their codons when aligned without reference to the amino acid alignment. Future research will explore the effect on PID of different alignment strategies.

Despite the differences in the PIDs for the random and shuffled sequence sets, the overall trend is that both G+C content and sequence length has an effect on sequence similarity (as measured by PID) (Fig. 3.11 & 3.12). This was further explored using the 50 pairs subset sequence pairs to see how the twilight values shifted for each alignment algorithm with respect to G+C content and sequence length.

4.2 Confounding factors

During the course of this research confounding factors were identified that could have been taken into account while designing the selection procedure for the PID subsets. G+C content, sequence length and the balance of genes for each PID all have the ability to affect the twilight zone (Fig. 3.9 & 3.10).

4.2.1 G+C content and length

The G+C content and the length of sequence pairs in the 50 pairs subset caused variation in the twilight value for each of the four core alignment algorithms tested (Fig. 3.9 & 3.10). The overall G+C content trend for all alignment algorithms was that pairs of sequences where both have high or low G+C content (G0-G0 or G2-G2) cause the twilight value to shift lower, and pairs of sequences where one has high and the other has low G+C content (G1-G2) cause the twilight value to shift slightly higher (Fig. 3.9). Sequence length also had a shared trend among all four alignment algorithms. Short or medium length sequence pairs shifted the twilight value higher whereas long length sequence pairs shifted the twilight value lower (Fig. 3.10).

G+C content

Sequence pairs with one high and one low G+C content sequence (G0-G2) shifted the twilight value 1% higher than the overall twilight value for `ssearch36`, `nhmmer` and `ggsearch` (Fig. 3.9). The twilight value decreased by 5% for pairs of sequences with the same high or low G+C content (G0-G0, G2-G2) for all alignment algorithms (Fig. 3.9). Therefore, having the same high or low G+C content has more effect on `ssearch36`, `nhmmer` and `ggsearch` than having opposite G+C content for this data set. The twilight value for `BLASTn` was 4% higher for sequence pairs with one high and one low G+C content which is a larger increase than all the other alignment algorithms (1%)(Fig. 3.9). There was a dip in the trend around the twilight value for `BLASTn` that may be due to the small number of pairs for each group rather than the effect of G+C content on sensitivity (Fig. 3.9). Using a larger number of pairs in future analyses will help determine if `BLASTn` is as equally effected by sequence pairs with one high and one low G+C content and pairs with the same high or low G+C content.

The effect of G+C content on the twilight zone is due to alignment algorithms maximising for similarity when aligning a pair of sequences. If sequences have similar high or low G+C content (G0-G0 or G2-G2) they will appear more similar then sequences with a large difference in G+C content, as shown by the PID distributions for the random and shuffled sets (Fig. 3.11 & 3.12). The PID of each sequence pair was calculated as measurement of divergence between sequences from a reference alignment determined by the CM or HMM for each gene (Fig. A.3) (*74*, *75*). When the homology search algorithms were used to search the database they used one query sequence (`nhmmer` HMMs are made from a single sequence query) and aligned the query to database sequences maximising for sequence similarity. Therefore, sequence pairs that both have the same high or low G+C content (G0-G0 or G2-G2) can appear to be more similar, then they were based on the alignment to the HMM or CM, due to their skewed nucleotide frequency. This means homology search algorithms are more likely to assign a more significant score to these sequences than a sequence pair of the same PID that has less similar G+C content. Therefore, the bit scores for non-homologous database sequences will be more similar to homolog pairs where one has high and the other low G+C content than homologs with the same high or low G+C content. This means the G+C content group where pairs have one high and one low G+C content with have a lower sensitivity. This causes the twilight zone to shift higher as these pairs reach a sensitivity of 0.5 for a higher PID.

Sequence length

The length of sequences in the subset also affected where the twilight value was for each alignment algorithm (Fig. 3.10 & 3.7). Sequence pairs with less then 1000 nucleotides shifted the twilight value higher (by a maximum of 3%) and those with 1000 nucleotides or more shifted the twilight value lower (by a maximum of 2%) for all alignment algorithms. This means that longer sequences have higher sensitivity for each PID value (on average) than the overall twilight value for each alignment algorithm (Fig. 3.10). They also have higher sensitivity than sequence pairs less than 1000 nucleotides long.

This is due to a similar phenomenon as G+C content has on alignment algorithms. The chance that a random sequence aligns to a query with a similar bit score to true homologs is dependent on the query length. Short query sequences have a smaller maximum bit score (query aligned to itself) then longer sequences, as there are fewer nucleotides to contribute to the bit score. Therefore, similarities between the query and the non-homologous sequence by chance, appear more significant for short queries as they are larger proportional to the maximum bit score possible and thus more similar to true homologs bit scores (Fig. 4.3). Sensitivity is calculated based on the number of true homologs found above the similarity bit score threshold set for a fixed FPR. Thus, short genes will have fewer true homologs included above the threshold compared with longer genes for the same PID value, due to the similar bit scores of non-homolgous and homologous sequences. This is why short sequence pairs had higher twilight values than long sequence pairs (Fig. 3.10). This is true for both global and local alignment algorithms, however local algorithms have the potential to have a higher bit score for non-homologous sequences than global as only locally conserved regions are scored (*4*, *5*). This allows for only the similar part of the query and the non-homolog to be aligned, rather than the whole length. This may explain why the variation in the twilight value due to length is smallest for `ggsearch` (Fig. 3.10)

The combined effect of G+C content and sequence length

Each sequence length group contains pairs with different G+C content. As G+C content also impacts the location of the twilight value, an even representation of G+C content groups would be required to minimise the effect of G+C content bias on the twilight value for each sequence length group. When selecting sequences for each PID value, G+C content and length were not taken into consideration. The 50

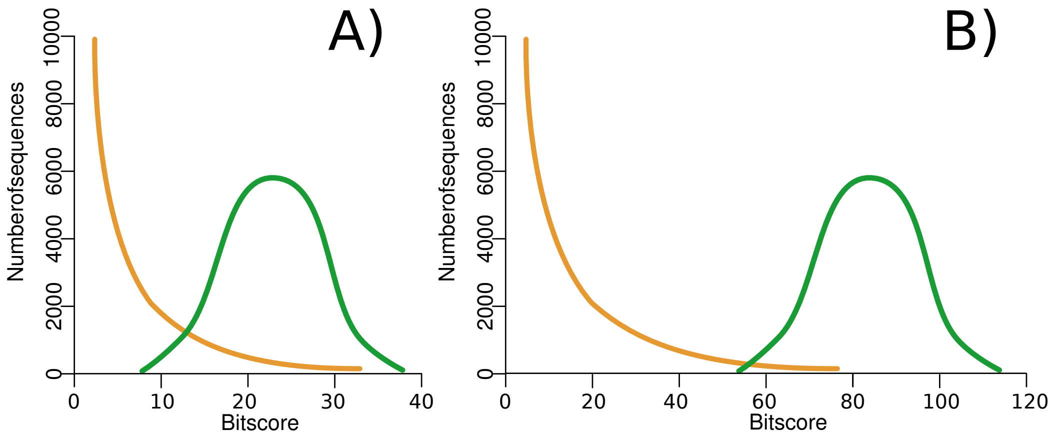


FIGURE 4.3: **Illustration of the affect of sequence length on local alignment algorithms:** *A) B)*
Green: True Homologs, Orange: Non-Homologs

pairs subset therefore does not have pairs for each sequence length and G+C content combination (Tab. 4.1). This is illustrated by the fact there is only one sequence pair with high G+C content for the short length sequence pairs with 40% PID (Tab. B.11). Therefore, the effects of G+C content and sequence length cannot be isolated from each other so the effect of each on the location of the twilight value cannot be independently tested for this data set.

	G0-G0	G0-G1	G0-G2	G1-G1	G1-G2	G2-G2	Total
L0-L0				524	91	150	**765**
L0-L1							**0**
L1-L1	52	179	176	602	344	229	**1,582**
L1-L2	2	15	29		51	21	**118**
L2-L2	37	255	433	17	528	371	**1,641**
Total	**91**	**449**	**638**	**1,146**	**1,014**	**771**	

TABLE 4.1: **Number of pairs in each G+C content and sequence length group:** Illustration of the current gaps in G+C content and sequence length groups (e.g. L0-L0+G0-G0) for the 50 pairs subset as well as the total number of sequences for each G+C content or sequence length group. The majority of pairs are of medium or long length with short sequences have fewer than half the number of medium or long sequences. The G+C content is more evenly spread although there are fewer pairs with low G+C content (G0-G0) then any other G+C content group. L0: Short sequence, L1: Medium length sequence, L2: Long length sequence. G0: Low G+C content, G1: Mid G+C content, G2: High G+C content

In order to determine the extent of each effect on the twilight value shift each G+C content group needs to be split into length groups for every PID value. There are 30 different G+C content and sequence length combinations made up of six G+C content and five sequence length groups (no short versus long sequence pairs) (Tab. 4.1). This would mean an average of 1.67 sequences for each combination per PID value. That means there are unlikely to be sufficient pairs if any, for each combination making it almost impossible to determine which set of factors is having an effect on the alignment algorithms for the 50 pairs PID subset. A quick check of the possible numbers per grouping for the 40% PID value shows that 18 different G+C content and sequence length combinations have no pairs and only five groupings have five pairs or more (10% of the subset) (Tab. B.11). This means the sensitivity calculated for each PID value based on the G+C content or sequence length groups is likely from an under powered sample of pairs. Therefore, the sparse nature of the data populating the G+C content and length groups likely does not capture the variation in the twilight value caused by G+C content and length. Hence the range of twilight values for different G+C content and sequence length groups where not included in the overall twilight zone reported (Fig. 3.13)(Tab. 3.3). If it was included this would have resulted in all alignment algorithms having a slightly larger range (+/- 3% at most) for their twilight zone (Fig. 4.4).

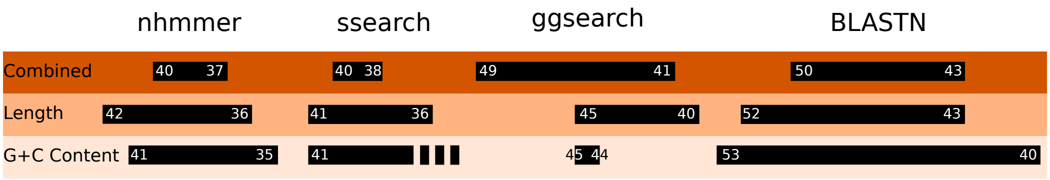


FIGURE 4.4: **G+C content and sequence length ranges for the twilight zone:** Each range shows the upper and lower bound for the twilight zone for each alignment algorithm. ***Combined*:** The range determined from the bootstrap data (Fig 3.8. ***Length*:** The range determined by the highest and lowest PID value from where each sequence length group (e.g. L0-L0) reached a sensitivity of 0.5 (Fig. 3.10). The upper bound (highest PID) is from the L0-L0 or L1-L1 group and the lower bound (lowest PID) is from the L2-L2 group. ***G+C Content*:** The range determined by the highest and lowest PID value from where each G+C content group (e.g. G0-G1) reached a sensitivity of 0.5 (Fig. 3.9). The upper bound is from the G0-G2 group and the lower bound if from the G0-G0 and G2-G2 groups combined.

The small number of pairs for each G+C content or sequence length group was also the reason smoothing splines were used to determine the trend for each group, rather then connected points like for the overall twilight values for each alignment algorithm (Fig. 3.9 & 3.10). This was required to smooth the effect of PID values where a group had only one pair available and so could only have a sensitivity of 1 or 0. This is why a few PID values between 40-70% have sensitivity values of 0 for both sequence length and G+C content groups (Fig. 3.9 & 3.10). There are also some sensitivities of 1 for PID values between 20-50% for the G+C content group with sequence pairs with the same high or low G+C content ((Fig. 3.9).

Ideally for future analyses, each PID value should have the same number of pairs for each G+C content and sequence length group to allow the comparison of the three groups with confidence that the effect seen is due to sequence composition.

A power analyses will be required to determine the minimum number of pairs to select that will allow detection of the effect of G+C content and sequence length. Then a mixed-effects logistic regression approach could be used to determine which sequence composition factor has the most effect on the twilight zone.

4.2.2 Representative gene set

In Gardner *et al.* 2007 it was shown that an unbalanced set of genes caused bias in the data set that led to a conclusion being drawn about the twilight zone of alignments that may be inaccurate (*11*, *64*). In the data set generated for this thesis, sequences were selected from a wide range of core genes and genera (Fig. 3.3)(Tbl. B.2 & B.3). This meant that most genes had sequence pairs from a wide range of PID values due to the wide taxonomic range of the bacteria chosen. This allowed a set of pairs for each PID value to be chosen that was not heavily weighted towards one gene (≤20% of pairs). Although this was not the case below 10% for both PID subsets due to the lack of genes in this subset (Tab. 3.1 & 3.2). These PID values were ignored for the overall twilight zone estimation due to the protein pairs likely being in the protein homology twilight zone (Fig. 3.4).

The database generated was an improvement on the four ncRNA gene database used by Gardner *et al.* and the 452 unique RNA chains from PDB used by Capriotti *et al.* 2010. However, knowledge gained during this thesis about factors that effect the range of each alignment algorithms twilight zone mean that further steps can be taken to improve the suitability of the database for calculating the twilight zone of nucleotide homology.

4.3 Future Work

The research carried out in this thesis has lead to further understanding of where the twilight zone of nucleotide homology lies. However it has also shown that there are some confounding factors that need to be factored into the experimental design in order to confidently assess the twilight zone.

4.3.1 True homolog set

There were two issues with determining the true homolog set for mRNA core genes that had to be resolved during the course of the analyses; ignoring nucleotide PIDs that correspond to the protein homology twilight zone and removal of duplicate sequences (Fig. 3.4)(Tab. B.9). Both of these problems could be resolved earlier in the analyses by the addition of processes during PID subset selection (e.g. removal of sequence pairs in the amino acid twilight zone) or the determination of core gene sequences (e.g. using protein structure).

Some of the sequence pairs included for the twilight zone analyses had amino acid PID values that were in the protein twilight zone for HMMER 3.0 (54). As nucleotide and amino acid PID were strongly correlated sequence pairs with nucleotide PID with ≤20% were ignored as they correspond to the amino acid twilight zone (Fig. 3.4). In future work sequence pairs that have amino acid PID values in the twilight zone for HMMER 3.0 will be removed from the pool of potential sequence pairs during PID subset selection (54). Sequences with PID values in the twilight zone could be true homologs but it is not possible to determine their homology with confidence based on sequence similarity, so they need to be excluded (52).

In the analyses performed in this thesis a set of protein sequences were found to have ambiguous homology as they were the top scoring sequence for a genome to two different genes (Tab. B.9). These sequences were all from bacteria in the Chloroflexi phylum and matched to multiple GTPase profile HMMS. All but three of the pairs of sequences had amino acid PIDs in or below the twilight zone (11%) (54). These sequences are either the homolog of one of the GTPases or a random high scoring match to the model. Although the easiest solution in the case of duplicate matches to genes is to remove the sequence, it does highlight a need to include more information than amino acid sequence to determine the true homologs. There are two possible improvements to the current method that would increase the information used to infer homology; including structural similarity or utilising Pfam and EggNOG HMMs to infer the homology of proteins.

SCOP

The structural classification of proteins database (SCOP) is a database of proteins collated into families based on their structure and amino acid sequence (103, 104). Each peptide chain (individual chain of amino acids) in a protein is considered as a protein domain and is classified into families based on structure. Using the sequences in protein families from SCOP for our protein core gene sequences, rather than the current approach, would give a higher level of evidence for calling sequences true homologs as the inference would be based on structural and sequence similarity (Tab. 4.2)(104). However, SCOP may not have sequence representatives of some of the core protein genes and there may not be a sequence for each genome from the representative subset due to all sequences in SCOP originating in PDB (56, 104). Therefore, the proteins available are dependent on which proteins are able to be structurally characterised experimentally as well as which proteins researchers are interested in.

Method	Advantages
SCOP	Sequences are assigned homology based on shared structural and sequence similarities
Pairmark	Simulates sequences found in databases (i.e. full contigs)
EggNOG + Pfam HMMs + Rfam CMs	Allows own choice of genomes and for a wide taxonomic range to be utilised. EggNOG and Pfam HMM combination allows both the whole sequence and conserved domains to inform homology assignment for the core genes.
Parametric Block Design	Allows better control of sequence confounders such as G+C content, sequence length and disproportionate gene representation
Simulated Evolution (e.g. Rose)	Can determine the G+C content and sequence length so each PID value has a lot of sequence pairs to choose from. Alignment accurately describes evolutionary events.

TABLE 4.2: **Method improvement summary:** Potential method improvements and the associated advantages and disadvantages.

EggNOG and Pfam HMMs

A more rigorous approach to using HMMs could also be adopted. Pfam HMMs are manually curated to include a gathering threshold that separates true and false homologs as best as is possible (15). Unlike EggNOG HMMs, Pfam HMMs are not always for the whole peptide chain sequence, instead they are probabilistic models for protein domains (15, 83). Protein domains are segments of a protein that are functionally and or structurally distinct from the rest of the protein sequence and usually play a functional role or facilitate interactions with other substrates or proteins (*15*). One protein sequence can have multiple copies of the same domain and one domain can be in many different proteins. Proteins contain highly conserved areas (e.g. functional domains) and variable regions in which the sequence of amino acids is not conserved as much. Pfam HMMs could be used in conjunction with the EggNOG HMMS to find sequences that have a significant bit score for the whole gene sequence and for conserved functional domains. This would have potentially circumvented the GTPase duplicate sequence problem as highly conserved regions were lacking in some of these sequences (Tab. C.1 Link C). As HMMs can be used on any genome this approach is preferred to SCOP as there is more control over the taxonomic diversity of the sequences and the genes to be included in the database.

4.3.2 Parametric block design

The impact seen on the nucleotide homology twilight value due to G+C content and sequence length requires a more stringent approach to the selection of pairs for each PID value than the current block design. At present, the only stipulation on each PID value subset is that each gene can only be up to 20% of the homolog pairs. An even distribution of G+C content and sequence lengths for each PID value will balance out any bias caused by either, and allow for exploration of the effect of each on the twilight zone. This will require relatively large numbers of pairs from each G+C content and sequence length group to prevent the sensitivity calculation for each group being strongly affected by small changes in the number of sequence pairs above the similarity threshold.

The range of G+C content and sequence lengths in the core gene set will be roughly cut into thirds to facilitate this whilst allowing a large enough set of sequences to populate each group. The G+C groups will remain the same, less than 40% (G0), 40-50% (G1) and greater than 50% (G2) as they already roughly cut the database and the known range of PID for all bacteria (13-75%) into three (Fig. 3.2) (*105*). The sequence length groups will change to allow for more even numbers of pairs in each length category. The lower quartile and upper quartile of core gene set will be used (rounded to a sensible number) to group the sequences into lengths that roughly divide the data into three. The short group will have sequences of less than 400 nucleotides in length (L0), the medium length group will have sequences between 400 - 1500 nucleotides (L1) and the long group will be sequence of 1500 nucleotides or more (L2).

This will make six possible combinations of G+C content and five possible combinations of sequence length for sequence pairs. It is not possible to have a sequence pair with one short and one long sequence as the thresholds we impose to determine homology prevent this combination from being possible. This means there will be 30 possible G+C content and sequence length combinations for each PID value. To discern the number of pairs required for each G+C content and length group a power analysis to determine the minimum sample size would need to be performed (106). This is to prevent the same problems that occurred for the length and G+C content division of the 50 pairs subset due to under powered sample sizes (Fig. 3.9 & 3.10) (107).

For each G+C content and sequence length group for each PID value the pairs will be as evenly split between ncRNA and mRNA as possible whilst maintaining ≤20% of pairs per gene. There are likely to be some PID values where there aren't enough ncRNA pairs due to the small number (9) of core ncRNA genes pairs are selected from (Tab. B.2). More ncRNA genes could be added, although they need to have a CM available. A quick search of the Rfam database shows other Rfam CMs with similar numbers of species (more than 2,000 species) to some of the core genes used in this research (74, 108). These CMs would need to be checked to see if the gene is vertically inherited and found in a wide phylogenetic range of species like the ncRNA core genes already selected.

A potentially less time consuming method to find pairs of sequences to fill gaps in PID value groups would be to simulate evolution using ROSE, or a similar tool, to create sets of 'homologous' sequences (109). ROSE allows for a sequence of specified length to be given as a starting point as well as the frequency of each nucleotide for insertions and substitutions to be set which would allow for G+C content to be increased or decreased compared with the ancestral sequence (109). This would allow us to control the sequence set to include sequences for specific PIDs, G+C contents and sequence lengths. The level of control this method affords the researcher may make finding sequence pairs to fill gaps faster than trying multiple new ncRNA CMs in the hope that the sequences for each genome fill the gaps. ROSE also produces an alignment based on the mutations that took place during the simulated evolution that could be used as the reference alignment for calculating PIDs.

The improved method described above for the selection of pairs based on sequence composition for each PID, as well as gene type, will give a more balanced and diverse data set to perform twilight zone analyses on.

4.3.3 A more realistic sequence database design

The majority of nucleotide sequences in databases available online have large sequences such as whole genomes, chromosomes or contigs, not individual gene sequences like the database utilised in this thesis. Therefore, embedding true homologs and shuffled variants into long, randomly generated contigs for future analyses will allow for a similar database design to those used in other analyses (Tab. 2.1) (54, 59, 110).

4.4 Conclusions

From this research it can be concluded that, whilst this most recent calculation of the nucleotide homology twilight zone has shown that the twilight zone is lower than expected, more work is required to further enhance the balance of the database. In particular this is with respect to balancing G+C content and sequence length in the PID subset.

The research detailed in this thesis has addressed the three original aims. The first aim was to assemble a representative sample of RNA sequences to create a balanced database. The database created contains a taxonomically diverse set of sequence pairs from core gene sequences that range from 0-100% PID (Fig. 3.1 & 3.3). The next aim was to observe any difference between old and current versions of the alignment algorithms. In this instance there was no difference between the old and new version of `ssearch` in terms of the twilight value although `ssearch34` is much slower than `ssearch36` (Fig. 3.7)(Tab. 3.4). The final aim was to identify the twilight zone of nucleotide homology which was found to be between 37-50% (Fig. 3.13). This means when researchers are using homology search algorithms to infer homology between nucleotide sequences they should only do so if the PID of the sequence pair is above 50%. If the PID is in the twilight range, they are required to perform further analyses to determine if the sequence pair is homologous or not (e.g. determining structural similarity, genomic context, shared motifs or domains, and functional assays).

Bibliography

(*1*) Loman, N. J. and Pallen, M. J. "Twenty years of bacterial genome sequencing". *Nature Reviews Microbiology* 13.12 (Dec. 2015), pp. 787–794.

(*2*) Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., and Ussery, D. W. "Insights from 20 years of bacterial genome sequencing". *Functional & Integrative Genomics* 15.2 (2015), pp. 141–161.

(*3*) Mount, D. W. and Mount, D. W. *Bioinformatics: sequence and genome analysis*. Vol. 2. Cold spring harbor laboratory press New York: 2001.

(*4*) Smith, T. and Waterman, M. "Identification of common molecular subsequences". *Journal of Molecular Biology* 147.1 (Mar. 1981), pp. 195–197.

(*5*) Needleman, S. B. and Wunsch, C. D. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of molecular biology* 48.3 (1970), pp. 443–453.

(*6*) Higgins, D. G. and Sharp, P. M. "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer". *Gene* 73.1 (Dec. 15, 1988), pp. 237–244.

(*7*) Saitou, N. and Nei, M. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution* 4.4 (July 1, 1987), pp. 406–425.

(*8*) Pearson, W. R. "Rapid and sensitive sequence comparison with FASTP and FASTA". *Methods in Enzymology* 183 (Jan. 1, 1990), pp. 63–98.

(*9*) BLAST, G. "PSI-BLAST: a new generation of protein database search programs Altschul". *Stephen F* (1997), pp. 3389–3402.

(*10*) Eddy, S. R. "Hidden markov models". *Current opinion in structural biology* 6.3 (1996), pp. 361–365.

(*11*) Gardner, P. P., Wilm, A., and Washietl, S. "A benchmark of multiple sequence alignment programs upon structural RNAs". *Nucleic Acids Research* 33.8 (Jan. 1, 2005), pp. 2433–2439.

(*12*) Bader, D. A. "Computational biology and high-performance computing". *Communications of the ACM* 47.11 (2004), pp. 34–41.

(*13*) Goodman, N. "Biological data becomes computer literate: new advances in bioinformatics". *Current Opinion in Biotechnology* 13.1 (Feb. 2002), pp. 68–71.

(*14*) Ronquist, F., Teslenko, M., Mark, P. van der, Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. "MrBayes 3.2:

Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space". *Systematic Biology* 61.3 (May 1, 2012), pp. 539–542.

(*15*) Sonnhammer, E. L. L., Eddy, S. R., and Durbin, R. "Pfam: A comprehensive database of protein domain families based on seed alignments". *Proteins: Structure, Function, and Bioinformatics* 28.3 (1997), pp. 405–420.

(*16*) Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. "Rfam: an RNA family database". *Nucleic Acids Research* 31.1 (Jan. 1, 2003), pp. 439–441.

(*17*) Owen, R. "On the structure and homologies of the cephalic tentacles in the pearly nautilus". *Annals and Magazine of Natural History* 12 (1843), pp. 305–311.

(*18*) Darwin, C. and Wallace, A. R. "Evolution by natural selection." *Evolution by natural selection.* (1958).

(*19*) Doolittle, R. F. "Similar amino acid sequences: chance or common ancestry?" *Science (New York, N.Y.)* 214.4517 (Oct. 9, 1981), pp. 149–159.

(*20*) Pearson, W. R. "An Introduction to Sequence Similarity ("Homology") Searching". *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 0 3 (June 2013).

(*21*) Fitch, W. M. "Homology: a personal view on some of the problems". *Trends in Genetics* 16.5 (May 1, 2000), pp. 227–231.

(*22*) Parker, J., Tsagkogeorga, G., Cotton, J. A., Liu, Y., Provero, P., Stupka, E., and Rossiter, S. J. "Genome-wide signatures of convergent evolution in echolocating mammals". *Nature* 502.7470 (Oct. 2013), pp. 228–231.

(*23*) Bankoff, R. J., Jerjos, M., Hohman, B., Lauterbur, M. E., Kistler, L., and Perry, G. H. "Testing Convergent Evolution in Auditory Processing Genes between Echolocating Mammals and the Aye-Aye, a Percussive-Foraging Primate". *Genome Biology and Evolution* 9.7 (July 26, 2017), pp. 1978–1989.

(*24*) Fitch, W. M. "Distinguishing homologous from analogous proteins". *Systematic zoology* 19.2 (1970), pp. 99–113.

(*25*) Hall, B. K. *Homology: The Hierarchial Basis of Comparative Biology*. Google-Books-ID: 8SGvMlyMXHUC. Academic Press, Dec. 2, 2012. 500 pp.

(*26*) Jensen, R. A. "Orthologs and paralogs-we need to get it right". *Genome biology* 2.8 (2001), interactions1002. 1.

(*27*) Dufayard, J.-F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., and Perrière, G. "Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases". *Bioinformatics* 21.11 (June 1, 2005), pp. 2596–2603.

(*28*) Koonin, E. V. "Orthologs, Paralogs, and Evolutionary Genomics". *Annual Review of Genetics* 39.1 (2005), pp. 309–338.

(*29*) Doolittle, R. F. *Of Urfs And Orfs: A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, 1986. 118 pp.

(*30*) Pearson, W. R. and Lipman, D. J. “Improved tools for biological sequence comparison”. *Proceedings of the National Academy of Sciences of the United States of America* 85.8 (Apr. 1988), pp. 2444–2448.

(*31*) Eddy, S. R. “Profile hidden Markov models.” *Bioinformatics (Oxford, England)* 14.9 (1998), pp. 755–763.

(*32*) Thompson, J. D., Plewniak, F., and Poch, O. “A comprehensive comparison of multiple sequence alignment programs”. *Nucleic Acids Research* 27.13 (July 1, 1999), pp. 2682–2690.

(*33*) Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

(*34*) Higgs, P. G. and Attwood, T. K. *Bioinformatics and molecular evolution*. John Wiley & Sons, 2013.

(*35*) Wheeler, W. “Alignment characters, dynamic programming and heuristic solutions”. *Molecular Approaches to Ecology and Evolution*. Ed. by R. DeSalle and B. Schierwater. Basel: Birkhäuser Basel, 1998, pp. 243–251.

(*36*) Wheeler, T. J. and Eddy, S. R. “nhmmer: DNA homology search with profile HMMs”. *Bioinformatics* 29.19 (Oct. 1, 2013), pp. 2487–2489.

(*37*) Pearson, W. R. “Finding Protein and Nucleotide Similarities with FASTA”. *Current protocols in bioinformatics* 53 (Mar. 24, 2016), pp. 3.9.1–3.925.

(*38*) Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. “BLAST+: architecture and applications”. *BMC Bioinformatics* 10.1 (Dec. 15, 2009), p. 421.

(*39*) Pearson, W. R. “FASTA Search Programs”. *eLS* (2001).

(*40*) Farrar, M. “Striped Smith–Waterman speeds database searches six times over other SIMD implementations”. *Bioinformatics* 23.2 (Jan. 15, 2007), pp. 156–161.

(*41*) Camacho, C., Madden, T. L., Tao, T., and Morgulis, A. *BLAST® Command Line Applications User Manual*. National Center for Biotechnology Information (US), 2008.

(*42*) Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. *Nucleic Acids Research* 25.17 (Sept. 1, 1997), pp. 3389–3402.

(*43*) Eddy, S. R. “Accelerated profile HMM searches”. *PLoS computational biology* 7.10 (2011), e1002195.

(*44*) Yoon, B.-J. “Hidden Markov Models and their Applications in Biological Sequence Analysis”. *Current Genomics* 10.6 (Sept. 2009), pp. 402–415.

(*45*) Finn, R. D., Clements, J., and Eddy, S. R. “HMMER web server: interactive sequence similarity searching”. *Nucleic Acids Research* 39 (suppl_2 July 1, 2011), W29–W37.

(*46*) Eddy, S. R. and TJ, W. “HMMER User’s Guide” (Version 3.1b2 Feb. 2015), p. 221.

(*47*) Brenner, S. E. “Errors in genome annotation”. *Trends in Genetics* 15.4 (Apr. 1999), pp. 132–133.

(*48*) Smith, T. F. and Zhang, X. “The challenges of genome sequence annotation or “The devil is in the details””. *Nature Biotechnology* 15.12 (Nov. 1997), pp. 1222–1223.

(*49*) Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. “Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies”. *PLOS Computational Biology* 5.12 (Dec. 11, 2009), e1000605.

(*50*) Gonzalez, A., Vázquez-Baeza, Y., Pettengill, J. B., Ottesen, A., McDonald, D., and Knight, R. “Avoiding Pandemic Fears in the Subway and Conquering the Platypus”. *mSystems* 1.3 (June 28, 2016), e00050–16.

(*51*) Hsu, T., Joice, R., Vallarino, J., Abu-Ali, G., Hartmann, E. M., Shafquat, A., DuLong, C., Baranowski, C., Gevers, D., Green, J. L., Morgan, X. C., Spengler, J. D., and Huttenhower, C. “Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment”. *mSystems* 1.3 (June 28, 2016), e00018–16.

(*52*) Rost, B. “Twilight zone of protein sequence alignments”. *Protein engineering* 12.2 (1999), pp. 85–94.

(*53*) Capriotti, E. and Marti-Renom, M. A. “Quantifying the relationship between sequence and three-dimensional structure conservation in RNA”. *BMC bioinformatics* 11 (June 15, 2010), p. 322.

(*54*) Eddy, S. R. *Twilight Zone of Protein Homology*. May 2012.

(*55*) Sander, C. and Schneider, R. “Database of homology-derived protein structures and the structural meaning of sequence alignment”. *Proteins: Structure, Function, and Bioinformatics* 9.1 (Jan. 1, 1991), pp. 56–68.

(*56*) Bank, P. D. “Protein data bank”. *Nature New Biol* 233 (1971), p. 223.

(*57*) Schneider, R. *MaxHom*. Univeristy of Heidelberg, 1994.

(*58*) Nawrocki, E. P. and Eddy, S. R. “Query-Dependent Banding (QDB) for Faster RNA Similarity Searches”. *PLOS Computational Biology* 3.3 (Mar. 30, 2007), e56.

(*59*) Eddy, S. R. *HMMER: biological sequence analysis using profile HMMs: profmark benchmark readme*. original-date: 2015-08-30. Mar. 14, 2019. URL: https://github.com/EddyRivasLab/hmmer/blob/master/profmark/00README (visited on 03/24/2019).

(*60*) Sanger, F. and Coulson, A. R. “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. *Journal of Molecular Biology* 94.3 (May 25, 1975), pp. 441–448.

(*61*) Eddy, S. R. and Durbin, R. “RNA sequence analysis using covariance models”. *Nucleic Acids Research* 22.11 (June 11, 1994), pp. 2079–2088.

(*62*) Wilm, A., Mainz, I., and Steger, G. “An enhanced RNA alignment benchmark for sequence alignment programs”. *Algorithms for Molecular Biology* 1 (Oct. 24, 2006), p. 19.

(*63*) Thompson, J. D., Plewniak, F., and Poch, O. "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs." *Bioinformatics* 15.1 (Jan. 1, 1999), pp. 87–88.

(*64*) Löwes, B., Chauve, C., Ponty, Y., and Giegerich, R. "The BRaliBase dent—a tale of benchmark design and interpretation". *Briefings in Bioinformatics* 18.2 (Mar. 2017), pp. 306–311.

(*65*) Eddy, S. R. "Non-coding RNA genes and the modern RNA world". *Nature Reviews. Genetics* 2.12 (Dec. 2001), pp. 919–929.

(*66*) Ke, A. and Doudna, J. A. "Crystallization of RNA and RNA–protein complexes". *Methods*. Macromolecular Crystallization 34.3 (Nov. 1, 2004), pp. 408–414.

(*67*) Scott, W., Finch, J., and Klug, A. "The crystal structure of an AII-RNAhammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage". *Cell* 81.7 (1995), pp. 991–1002.

(*68*) Rife, J. P., Stallings, S. C., Correll, C. C., Dallas, A., Steitz, T. A., and Moore, P. B. "Comparison of the Crystal and Solution Structures of Two RNA Oligonucleotides". *Biophysical Journal* 76.1 (Jan. 1, 1999), pp. 65–75.

(*69*) Unknown. *NCBI RefSeq Database ftp Downloading instructions*. Dec. 25, 2016. URL: ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_Downloading_Genomic_Data.pdf (visited on 11/27/2017).

(*70*) Letunic, I. and Bork, P. "Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees". *Nucleic Acids Research* 44 (W1 2016), W242–245.

(*71*) May, A. C. "Percent Sequence Identity: The Need to Be Explicit". *Structure* 12.5 (May 2004), pp. 737–738.

(*72*) Umu, S. U., Poole, A. M., Dobson, R. C., and Gardner, P. P. "Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea". *eLife* 5 (Sept. 20, 2016).

(*73*) Gil, R., Silva, F. J., Peretó, J., and Moya, A. "Determination of the Core of a Minimal Bacterial Gene Set". *Microbiology and Molecular Biology Reviews* 68.3 (Sept. 2004), pp. 518–537.

(*74*) Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D., and Petrov, A. I. "Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families". *Nucleic acids research* 46 (D1 2017), pp. D335–D342.

(*75*) Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., and Kuhn, M. "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences". *Nucleic acids research* 44 (D1 2015), pp. D286–D293.

(*76*) Böck, A., Forchhammer, K., Heider, J., Leinfelder, W., Sawers, G., Veprek, B., and Zinoni, F. "Selenocysteine: the 21st amino acid". *Molecular Microbiology* 5.3 (1991), pp. 515–520.

(*77*) Zhang, Y., Romero, H., Salinas, G., and Gladyshev, V. N. "Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues". *Genome Biology* 7.10 (2006), R94.

(*78*) Lowe, T. M. and Eddy, S. R. "tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence". *Nucleic Acids Research* 25.5 (Mar. 1, 1997), pp. 955–964.

(*79*) Shortridge, M. D., Triplet, T., Revesz, P., Griep, M. A., and Powers, R. "Bacterial Protein Structures Reveal Phylum Dependent Divergence". *Computational biology and chemistry* 35.1 (Feb. 2011), pp. 24–33.

(*80*) Zarrineh, P., Sánchez-Rodríguez, A., Hosseinkhan, N., Narimani, Z., Marchal, K., and Masoudi-Nejad, A. "Genome-Scale Co-Expression Network Comparison across Escherichia coli and Salmonella enterica Serovar Typhimurium Reveals Significant Conservation at the Regulon Level of Local Regulators Despite Their Dissimilar Lifestyles". *PLOS ONE* 9.8 (Aug. 7, 2014), e102871.

(*81*) Hori, H. and Osawa, S. "Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences." *Molecular biology and evolution* 4.5 (1987), pp. 445–472.

(*82*) Kunisawa, T. "Identification and chromosomal distribution of DNA sequence segments conserved since divergence of Escherichia coli and Bacillus subtilis". *Journal of Molecular Evolution* 40.6 (June 1, 1995), pp. 585–593.

(*83*) Jensen, L. J., Julien, P., Kuhn, M., Mering, C. von, Muller, J., Doerks, T., and Bork, P. "eggNOG: automated construction and annotation of orthologous groups of genes". *Nucleic Acids Research* 36 (Database issue Jan. 2008), pp. D250–D254.

(*84*) Nawrocki, E. P. and Eddy, S. R. "Infernal 1.1: 100-fold faster RNA homology searches". *Bioinformatics* 29.22 (2013), pp. 2933–2935.

(*85*) Woese, C. R. "Bacterial evolution." *Microbiological Reviews* 51.2 (June 1987), pp. 221–271.

(*86*) Lerat, E., Daubin, V., and Moran, N. A. "From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria". *PLoS biology* 1.1 (Oct. 2003), E19.

(*87*) Nawrocki, E. P. "Structural RNA Homology Search and Alignment Using Covariance Models". PhD thesis. St. Louis, USA: Washington University, Jan. 2009.

(*88*) Albà, M. M. "Replicative DNA polymerases". *Genome Biology* 2.1 (2001), reviews3002.1–reviews3002.4.

(*89*) Eddy, S. R. "A new generation of homology search tools based on probabilistic inference". *Genome Informatics* 23 (2009). HMMER3, pp. 205–211.

(*90*) Suyama, M., Torrents, D., and Bork, P. "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments". *Nucleic Acids Research* 34 (Web Server issue July 1, 2006), W609–W612.

(*91*) Felsenstein, J. "Distance methods for inferring phylogenies: a justification". *Evolution* 38.1 (1984), pp. 16–24.

(*92*) Workman, C. and Krogh, A. "No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution". *Nucleic Acids Research* 27.24 (Dec. 1, 1999), pp. 4816–4822.

(*93*) Seffens, W. and Digby, D. "mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences". *Nucleic acids research* 27.7 (1999), pp. 1578–1584.

(*94*) Babak, T., Blencowe, B. J., and Hughes, T. R. "Considerations in the identification of functional RNA structural elements in genomic alignments". *BMC Bioinformatics* 8 (Jan. 30, 2007), p. 33.

(*95*) States, D, Gish, W, and Altschul, S. "Improved sensitivity of nucleic acid database searches using application-specific scoring matrices". *Methods* 3.1 (Aug. 1991), pp. 66–70.

(*96*) Felsenstein, J. "Confidence limits on phylogenies: an approach using the bootstrap". *Evolution* 39.4 (1985), pp. 783–791.

(*97*) Efron, B. and Tibshirani, R. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy". *Statistical science* (1986), pp. 54–75.

(*98*) Grunicke, H. H. and Maly, K. "Role of GTPases and GTPase regulatory proteins in oncogenesis". *Critical Reviews in Oncogenesis* 4.4 (1993), pp. 389–402.

(*99*) Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. "Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions". *Nucleic Acids Research* 41.12 (July 1, 2013), e121–e121.

(*100*) Scheffzek, K. and Ahmadian, M. R. "GTPase activating proteins: structural and functional insights 18 years after discovery". *Cellular and Molecular Life Sciences CMLS* 62.24 (Dec. 1, 2005), pp. 3014–3038.

(*101*) Spearman, C. "The Proof and Measurement of Association between Two Things". *The American Journal of Psychology* 15.1 (1904), pp. 72–101.

(*102*) Caballero, J., Smit, A. F. A., Hood, L., and Glusman, G. "Realistic artificial DNA sequences as negative controls for computational genomics". *Nucleic Acids Research* 42.12 (Aug. 1, 2014), e99.

(*103*) Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. "SCOP: A structural classification of proteins database for the investigation of sequences and structures". *Journal of Molecular Biology* 247.4 (Apr. 7, 1995), pp. 536–540.

(*104*) Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. "SCOP2 prototype: a new approach to protein structure mining". *Nucleic Acids Research* 42 (D1 Jan. 2014), pp. D310–D314.

(*105*) Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., and Daubin, V. "GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands". *PLOS Genetics* 11.2 (Feb. 6, 2015), e1004941.

(*106*) Cohen, J. "Statistical Power Analysis". *Current Directions in Psychological Science* 1.3 (1992), pp. 98–101.

(*107*) Brewer, J. K. "On the power of statistical tests in the American Educational Research Journal". *American Educational Research Journal* 9.3 (1972), pp. 391–401.

(*108*) Storz, G. "An expanding universe of noncoding RNAs". *Science (New York, N.Y.)* 296.5571 (May 17, 2002), pp. 1260–1263.

(*109*) Stoye, J., Evers, D., and Meyer, F. "Rose: generating sequence families". *Bioinformatics* 14.2 (Mar. 1, 1998), pp. 157–163.

(*110*) Eddy, S. R. *RMARK Readme: Infernal - RNA secondary structure/sequence profiles for homology search and alignment*. original-date: 2015-08-30. Mar. 1, 2019. URL: https://github.com/EddyRivasLab/infernal (visited on 03/25/2019).

(*111*) Kunisawa, T. "The phylogenetic placement of the non-phototrophic, Gram-positive thermophile 'Thermobaculum terrenum' and branching orders within the phylum 'Chloroflexi' inferred from gene order comparisons". *International Journal of Systematic and Evolutionary Microbiology* 61.8 (2011), pp. 1944–1953.

Appendix A

Supplementary Figures

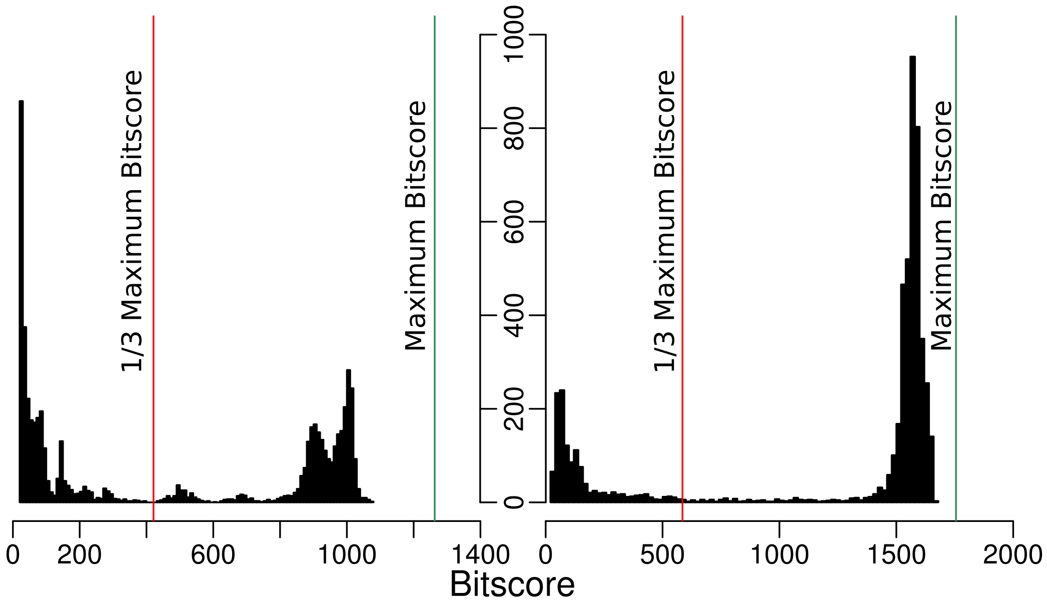


FIGURE A.1: **Distribution of bit scores of sequences in the NCBI RefSeq genomes to the ENOG4105C3H HMM and RF00177 CM:** Bit score distributions showing the characteristic bimodal distributions for HMM and CM alignments. The green line shows the maximum bit score possible, gained by emitting the consensus sequence for each HMM or CM and aligning it back to the model. The red line shows how 1/3 of the maximum bit score cuts off the lower mode distribution that is generally partial matches and noise. This same trend where 1/3 the maximum bit score removes the 'noise' mode distribution can be seen for all other models.
Left: ENOG4105C3H. *Right*: RF00177

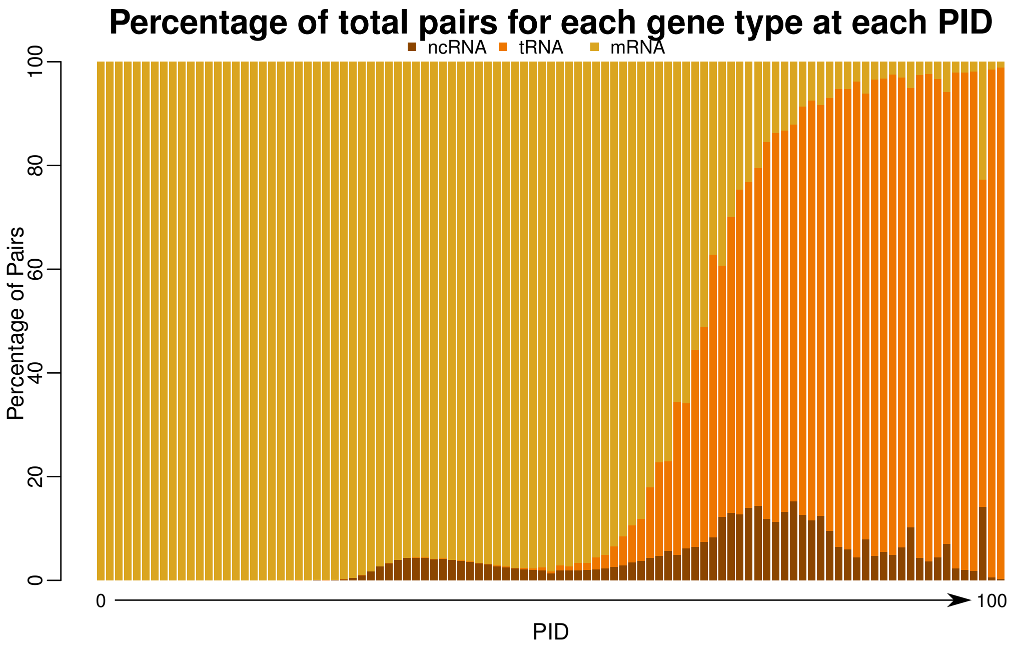


FIGURE A.2: **Proportion of gene type for each PID value for the core gene set:** This shows the balance of mRNA, tRNA and ncRNA for each PID value. tRNA is the majority of pairs for high PID values whereas mRNA is the majority of pairs for low PID values. ncRNA is always less then 20% of the pairs available per PID.

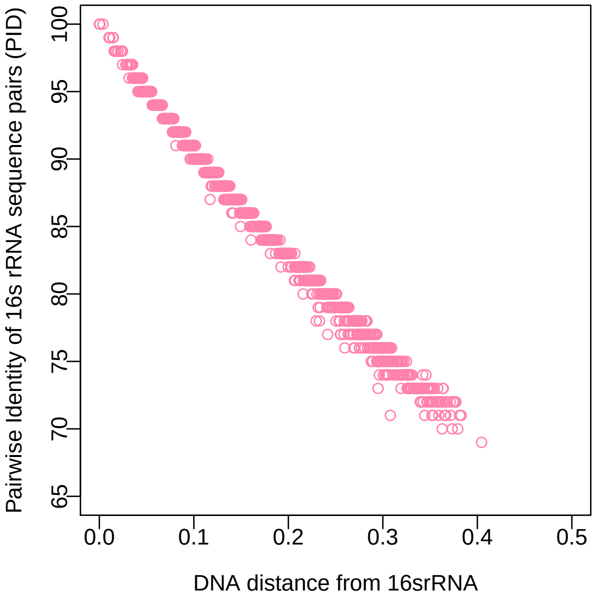


FIGURE A.3: **PID versus F84 phylogenetic measure for 16s rRNA sequence pairs:** The Pearson correlation coefficient between 16s rRNA PID and the F84 is -0.995 (three decimal places). This shows there is a strong negative correlation between the F84 DNA distance measure and PID.

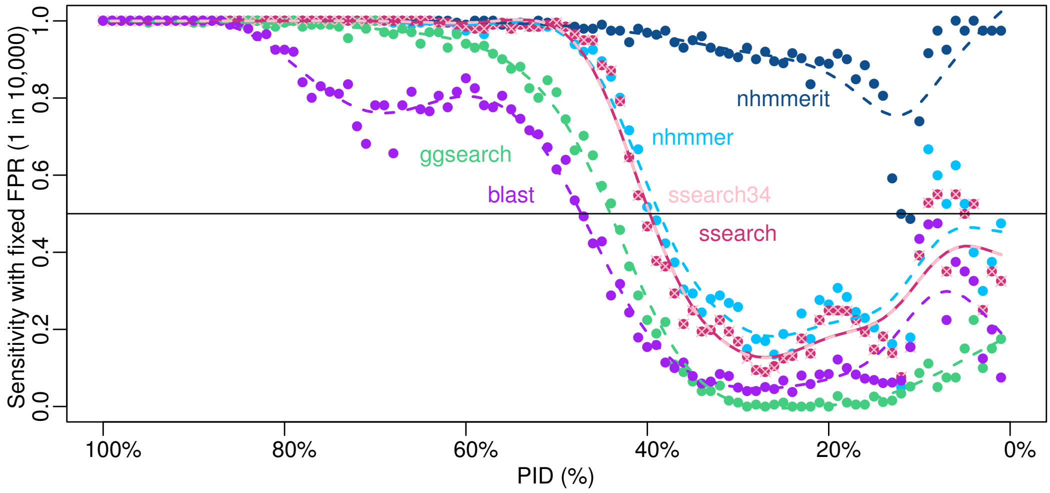


FIGURE A.4: **Twilight zone for the 201 pairs subset showing the full range of PID values:** This graph illustrates how the trend for sensitivity as PID decreases. Each trend line is a R smooth.spline with 12 degrees of freedom. Below 20% the sensitivity of the four core alignment algorithms and `ssearch34` increases again.

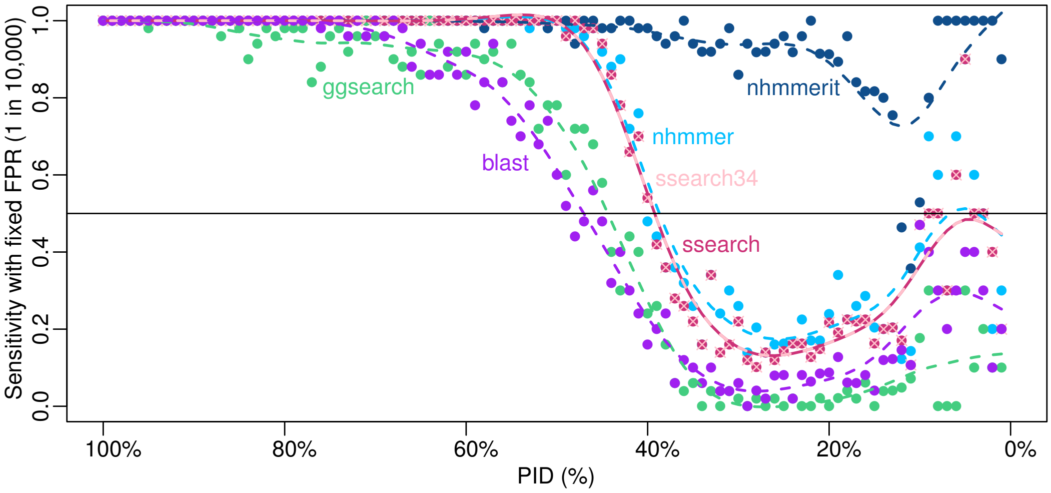


FIGURE A.5: **Twilight zone for the 50 pairs subset showing the full range of PID values:** This graph illustrates how the trend for sensitivity as PID decreases. Below 20% the sensitivity of the four core alignment algorithms and `ssearch34` increases again.

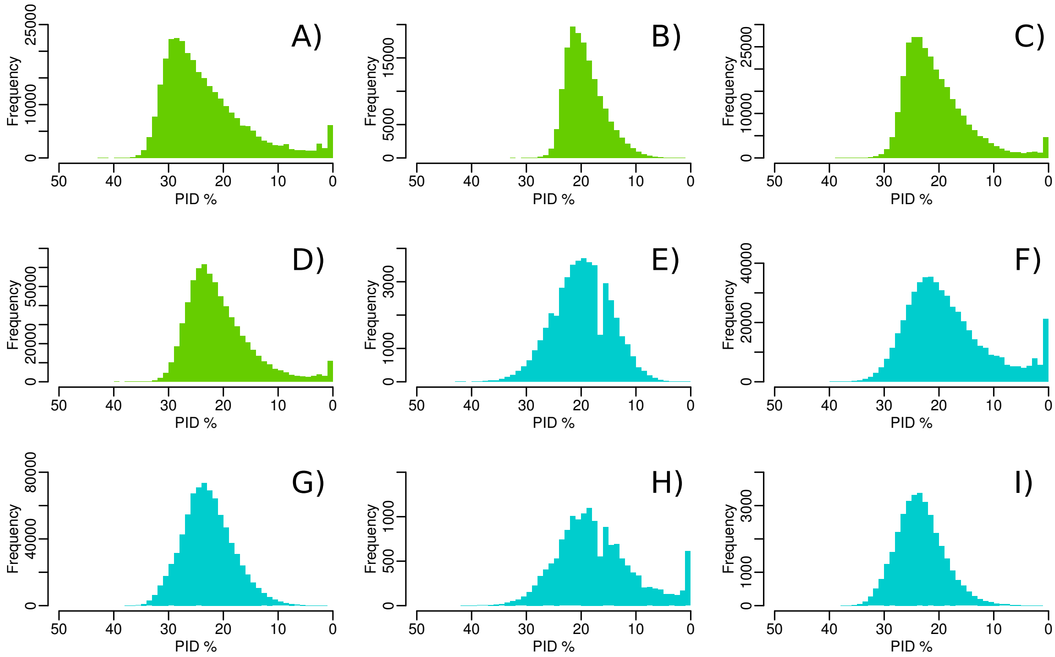


FIGURE A.6: **PID distribution for G+C content and sequence length groups for the shuffled sequence set: A)** Same GC content for both sequences either high or low (G2-G2 or G0-G0), mode = 29%, maximum value = 43% **B)** Opposite GC content for each sequence, high and low (G0-G2), mode = 22%, maximum value = 33% **C)** Medium G+C content sequence pairs (G1-G1), mode = 24%, maximum value = 40% **D)** G+C content where one sequence has medium and the other has high or low G+C content (G0-G1 or G1-G2), mode = 24%, maximum value = 41% **E)** Two short sequences (L0-L0), mode = 20%, maximum value = 43% **F)** Two medium length sequences (L1-L1), mode = 22%, maximum value = 42% **G)** Two long length sequences (L2-L2), mode = 24%, maximum value = 39% **H)** One short and one medium length sequence (L0-L1), mode = 19%, maximum value = 42% **I)** One medium and one long length sequence (L1-L2), mode = 24%, maximum value =38%

Appendix B

Supplementary Tables

Software	Citations of Original Paper	Citations of Newest Paper
Blast	76,015 (1990)	4,775 (2009 - BLAST+)
HMMER	583 (1995)	1,838 (2011 - HMMERv3)
FASTA	13,502 (1988)	70 (2016 - FASTA36)

TABLE B.1: **Number of citations for BLAST, HMMER and FASTA:** Number of Google Scholar citations for each method on the 10th Feburary at 1:30pm to the original paper (column 2) and the most recent full version release (column 3). All three suites of algorithms are widely used, as shown by the large number of combined citations.

TABLE B.2: **mRNA and tRNA core gene annotations:** The annotations for each CM from Rfam for each ncRNA gene. The tRNA annotations are expansions of the tRNA three letter codes used to identify the tRNAs by tRNAscan-SE.

Start of Table B.2

Gene Name	**Gene Annotation**
Ala	Alanine tRNA
Arg	Arginine tRNA
Asn	Asparagine tRNA
Asp	Aspartic acid tRNA
Cys	Cysteine tRNA
Gln	Glutamine tRNA
Glu	Glutamic acid tRNA
Gly	Glycine tRNA
His	Histidine tRNA
Ile	Isoleucine tRNA
Leu	Leucine tRNA
Lys	Lysine tRNA
Met	Methionine tRNA
Phe	Phenylalanine tRNA
Pro	Proline tRNA
SeC	Selenocysteine tRNA

Continuation of Table B.2	
Gene Name	**Gene Annotation**
Ser	Serine tRNA
Thr	Threonine tRNA
Trp	Tryptophan tRNA
Tyr	Tyrosine tRNA
Val	Valine tRNA
RF00001	5S Ribosomal RNA
RF00010	Bacterial RNase P class a
RF00011	Bacterial RNase P class b
RF00013	6S SsrS RNA
RF00023	Transfer messanger RNA (tmRNA)
RF00169	Bacterial small signal recognition particle RNA (small SRP)
RF00177	Bacterial small subunit RNA (16s Ribosomal RNA)
RF01854	Bacterial large signal recognition particle RNA (large SRP)
RF02541	Bacterial large subunit RNA (23s Ribosomal RNA)
End of Table B.2	

TABLE B.3: **mRNA EggNOG HMM gene annotations:** The annotations of the core gene HMMs that were selected from EggNOG to be used to find sequences within the genus representative bacterial genome set. Annotations were extracted from accompanying bactNOG annotations file from EggNOG. Bold HMM names are genes used in the DNA replication genome completeness test (6 hmms).

Start of Table B.3	
Eggnog ID	**Annotation**
ENOG4105BZ3	General (non sugar-specific) component of the phosphoenolpyruvate-dependent sugar phosphotransferase system (sugar PTS). This major carbohydrate active-transport system catalyzes the phosphorylation of incoming sugar substrates concomitantly with their translocation across the cell membrane. Enzyme I transfers the phosphoryl group from phosphoenolpyruvate (PEP) to the phosphoryl carrier protein (HPr) (By similarity)
ENOG4105BZ5	ATP binding to DnaK triggers the release of the substrate protein, thus completing the reaction cycle. Several rounds of ATP-dependent interactions between DnaJ, DnaK and GrpE are required for fully efficient folding. Also involved, together with DnaK and GrpE, in the DNA replication of plasmids through activation of initiation proteins (By similarity)

Continuation of Table B.3	
Eggnog ID	**Annotation**
ENOG4105BZ6	Presumably involved in the processing and regular turnover of intracellular proteins. Catalyzes the removal of unsubstituted N-terminal amino acids from various peptides (By similarity)
ENOG4105BZH	Provides the precursors necessary for DNA synthesis. Catalyzes the biosynthesis of deoxyribonucleotides from the corresponding ribonucleotides (By similarity)
ENOG4105C0B	DNA polymerase III (alpha subunit)
ENOG4105C0V	Provides the sole de novo source of dTMP for DNA biosynthesis (By similarity)
ENOG4105C17	glyceraldehyde-3-phosphate dehydrogenase
ENOG4105C1H	Exhibits a very high intrinsic GTPase hydrolysis rate. Involved in the addition of a carboxymethylaminomethyl (cmnm) group at the wobble position (U34) of certain tRNAs, forming tRNA- cmnm(5)s(2)U34 (By similarity)
ENOG4105C20	Catalyzes the attachment of glutamate to tRNA(Glu) in a two-step reaction glutamate is first activated by ATP to form Glu-AMP and then transferred to the acceptor end of tRNA(Glu) (By similarity)
ENOG4105C22	threonyL-tRNA synthetase
ENOG4105C24	DNA gyrase negatively supercoils closed circular double-stranded DNA in an ATP-dependent manner and also catalyzes the interconversion of other topological isomers of double-stranded DNA rings, including catenanes and knotted rings (By similarity)
ENOG4105C2M	dna polymerase I
ENOG4105C31	Tryptophanyl-tRNA synthetase
ENOG4105C38	Glycyl-tRNA synthetase beta subunit
ENOG4105C3G	gtp-binding protein
ENOG4105C3H	Acts as a processive, ATP-dependent zinc metallopeptidase for both cytoplasmic and membrane proteins. Plays a role in the quality control of integral membrane proteins (By similarity)
ENOG4105C3J	Cysteine desulfurase
ENOG4105C3M	thioredoxin reductase
ENOG4105C4J	Produces ATP from ADP in the presence of a proton gradient across the membrane. The catalytic sites are hosted primarily by the beta subunits (By similarity)

Continuation of Table B.3	
Eggnog ID	**Annotation**
ENOG4105C4S	Required for accurate and efficient protein synthesis under certain stress conditions. May act as a fidelity factor of the translation reaction, by catalyzing a one-codon backward translocation of tRNAs on improperly translocated ribosomes. Back- translocation proceeds from a post-translocation (POST) complex to a pre-translocation (PRE) complex, thus giving elongation factor G a second chance to translocate the tRNAs correctly. Binds to ribosomes in a GTP-dependent manner (By similarity)
ENOG4105C5S	asparaginyl-tRNA synthetase
ENOG4105C5T	Phosphoribosyl pyrophosphate synthase
ENOG4105C62	Involved in mRNA degradation. Hydrolyzes single-stranded polyribonucleotides processively in the 3'- to 5'-direction (By similarity)
ENOG4105C64	Binds directly to 23S rRNA. The L1 stalk is quite mobile in the ribosome, and is involved in E site tRNA release (By similarity)
ENOG4105C65	Catalyzes the reversible interconversion of serine and glycine with tetrahydrofolate (THF) serving as the one-carbon carrier. This reaction serves as the major source of one-carbon groups required for the biosynthesis of purines, thymidylate, methionine, and other important biomolecules. Also exhibits THF- independent aldolase activity toward beta-hydroxyamino acids, producing glycine and aldehydes, via a retro-aldol mechanism (By similarity)
ENOG4105C6A	phenylalanyl-tRNA synthetase (beta subunit)
ENOG4105C6P	ATP-dependent serine protease that mediates the selective degradation of mutant and abnormal proteins as well as certain short-lived regulatory proteins. Required for cellular homeostasis and for survival from DNA damage and developmental changes induced by stress. Degrades polypeptides processively to yield small peptide fragments that are 5 to 10 amino acids long. Binds to DNA in a double-stranded, site-specific manner (By similarity)
ENOG4105C70	Catalyzes the reversible conversion of 2- phosphoglycerate into phosphoenolpyruvate. It is essential for the degradation of carbohydrates via glycolysis (By similarity)

Continuation of Table B.3

Eggnog ID	**Annotation**
ENOG4105C77	DNA ligase that catalyzes the formation of phosphodiester linkages between 5′-phosphoryl and 3′-hydroxyl groups in double-stranded DNA using NAD as a coenzyme and as the energy source for the reaction. It is essential for DNA replication and repair of damaged DNA (By similarity)
ENOG4105C7D	DNA gyrase negatively supercoils closed circular double-stranded DNA in an ATP-dependent manner and also catalyzes the interconversion of other topological isomers of double-stranded DNA rings, including catenanes and knotted rings (By similarity)
ENOG4105C8D	Catalyzes the ATP-dependent amination of UTP to CTP with either L-glutamine or ammonia as the source of nitrogen (By similarity)
ENOG4105C8K	Peptide chain release factor 1 directs the termination of translation in response to the peptide chain termination codons UAG and UAA (By similarity)
ENOG4105C8N	cysteinyl-tRNA synthetase
ENOG4105C8T	Leucyl-tRNA synthetase
ENOG4105C90	Catalyzes the attachment of proline to tRNA(Pro) in a two-step reaction proline is first activated by ATP to form Pro-AMP and then transferred to the acceptor end of tRNA(Pro). As ProRS can inadvertently accommodate and process non-cognate amino acids such as alanine and cysteine, to avoid such errors it has two additional distinct editing activities against alanine. One activity is designated as ′pretransfer′ editing and involves the tRNA(Pro)-independent hydrolysis of activated Ala-AMP. The other activity is designated ′post-transfer′ editing and involves deacylation of mischarged Ala-tRNA(Pro). The misacylated Cys- tRNA(Pro) is not edited by ProRS (By similarity)
ENOG4105C9G	DNA primase is the polymerase that synthesizes small RNA primers for the Okazaki fragments on both template strands at replication forks during chromosomal DNA synthesis (By similarity)
ENOG4105C9M	aspartyl-trna synthetase

Continuation of Table B.3

Eggnog ID	Annotation
ENOG4105C9R	An essential GTPase which binds GTP, GDP and possibly (p)ppGpp with moderate affinity, with high nucleotide exchange rates and a fairly low GTP hydrolysis rate (By similarity). It may play a role in control of the cell cycle, stress response, ribosome biogenesis and in those bacteria that undergo differentiation, in morphogenesis control
ENOG4105CA1	Removes the N-terminal methionine from nascent proteins (By similarity)
ENOG4105CA4	amino acids such as threonine, to avoid such errors, it has a posttransfer editing activity that hydrolyzes mischarged Thr-tRNA(Val) in a tRNA-dependent manner (By similarity)
ENOG4105CA9	Pyruvate kinase
ENOG4105CB9	Involved in targeting and insertion of nascent membrane proteins into the cytoplasmic membrane. Binds to the hydrophobic signal sequence of the ribosome-nascent chain (RNC) as it emerges from the ribosomes. The SRP-RNC complex is then targeted to the cytoplasmic membrane where it interacts with the SRP receptor FtsY
ENOG4105CC8	Catalyzes the reversible transfer of the terminal phosphate group between ATP and AMP. Plays an important role in cellular energy homeostasis and in adenine nucleotide metabolism (By similarity)
ENOG4105CCJ	Catalyzes the 2-thiolation of uridine at the wobble position (U34) of tRNA, leading to the formation of s(2)U34 (By similarity)
ENOG4105CDG	Produces ATP from ADP in the presence of a proton gradient across the membrane. The alpha chain is a regulatory subunit (By similarity)
ENOG4105CDK	Essential cell division protein that forms a contractile ring structure (Z ring) at the future cell division site. The regulation of the ring assembly controls the timing and the location of cell division. One of the functions of the FtsZ ring is to recruit other cell division proteins to the septum to produce a new cell wall between the dividing cells. Binds GTP and shows GTPase activity (By similarity)
ENOG4105CDU	Replicative dna helicase
ENOG4105CE9	30S ribosomal protein S2

Continuation of Table B.3	
Eggnog ID	**Annotation**
ENOG4105CEJ	Catalyzes the GTP-dependent ribosomal translocation step during translation elongation. During this step, the ribosome changes from the pre-translocational (PRE) to the post-translocational (POST) state as the newly formed A-site-bound peptidyl-tRNA and P-site-bound deacylated tRNA move to the P and E sites, respectively. Catalyzes the coordinated movement of the two tRNA molecules, the mRNA and conformational changes in the ribosome (By similarity)
ENOG4105CEY	Amp-dependent synthetase and ligase
ENOG4105CFD	One of the primary rRNA binding proteins. Required for association of the 30S and 50S subunits to form the 70S ribosome, for tRNA binding and peptide bond formation. It has been suggested to have peptidyltransferase activity
ENOG4105CFG	Acts as a chaperone (By similarity)
ENOG4105CGG	The central subunit of the protein translocation channel SecYEG. Consists of two halves formed by TMs 1-5 and 6-10. These two domains form a lateral gate at the front which open onto the bilayer between TMs 2 and 7, and are clamped together by SecE at the back. The channel is closed by both a pore ring composed of hydrophobic SecY resides and a short helix (helix 2A) on the extracellular side of the membrane which forms a plug. The plug probably moves laterally to allow the channel to open. The ring and the pore may move independently (By similarity)
ENOG4105CGJ	Specifically methylates the N4 position of cytidine in position 1402 (C1402) of 16S rRNA (By similarity)
ENOG4105CGR	Catalyzes the attachment of serine to tRNA(Ser). Is also able to aminoacylate tRNA(Sec) with serine, to form the misacylated tRNA L-seryl-tRNA(Sec), which will be further converted into selenocysteinyl-tRNA(Sec) (By similarity)
ENOG4105CGV	This protein promotes the GTP-dependent binding of aminoacyl-tRNA to the A-site of ribosomes during protein biosynthesis (By similarity)
ENOG4105CHV	Transcription elongation factor NusA
ENOG4105CI6	Part of the Sec protein translocase complex. Interacts with the SecYEG preprotein conducting channel. Has a central role in coupling the hydrolysis of ATP to the transfer of proteins into and across the cell membrane, serving

Continuation of Table B.3	
Eggnog ID	**Annotation**
ENOG4105CIM	Catalyzes the attachment of alanine to tRNA(Ala) in a two-step reaction alanine is first activated by ATP to form Ala- AMP and then transferred to the acceptor end of tRNA(Ala). Also edits incorrectly charged Ser-tRNA(Ala) and Gly-tRNA(Ala) via its editing domain (By similarity)
ENOG4105CJ9	Prevents misfolding and promotes the refolding and proper assembly of unfolded polypeptides generated under stress conditions (By similarity)
ENOG4105CJS	Phosphopantothenoylcysteine decarboxylase
ENOG4105CKE	Binds the lower part of the 30S subunit head. Binds mRNA in the 70S ribosome, positioning it for translation (By similarity)
ENOG4105CKH	Is required not only for elongation of protein synthesis but also for the initiation of all mRNA translation through initiator tRNA(fMet) aminoacylation (By similarity)
ENOG4105CP7	Triose-phosphate isomerase
ENOG4105CPH	Catalyzes the formation of S-adenosylmethionine from methionine and ATP
ENOG4105CPM	Required for the formation of a threonylcarbamoyl group on adenosine at position 37 (t(6)A37) in tRNAs that read codons beginning with adenine (By similarity)
ENOG4105CRK	lysyL-tRNA synthetase
ENOG4105CSF	NADPH-dependent glycerol-3-phosphate dehydrogenase
ENOG4105CSM	endonuclease III
ENOG4105CSS	phenylalanyl-tRNA synthetase (alpha subunit)
ENOG4105CTF	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates (By similarity)
ENOG4105CU0	Catalyzes the 2'-O-methylation of the ribose of cytidine 1402 (C1402) in 16S rRNA (By similarity)
ENOG4105CU7	Associates with the EF-Tu.GDP complex and induces the exchange of GDP to GTP. It remains bound to the aminoacyl-tRNA.EF- Tu.GTP complex up to the GTP hydrolysis stage on the ribosome (By similarity)
ENOG4105CV1	Transketolase (EC 2.2.1.1)
ENOG4105CW6	This is 1 of the proteins that binds and probably mediates the attachment of the 5S RNA into the large ribosomal subunit, where it forms part of the central protuberance. In the 70S ribosome it contacts protein S13 of the 30S subunit (bridge B1b), connecting the 2 subunits

Continuation of Table B.3	
Eggnog ID	**Annotation**
ENOG4105CWT	An essential GTPase that binds both GDP and GTP, with rapid nucleotide exchange. Plays a role in 16S rRNA processing and 30S ribosomal subunit biogenesis and possibly also in cell cycle regulation and energy metabolism (By similarity)
ENOG4105CZ5	Catalyzes the conversion of uracil and 5-phospho-alpha- D-ribose 1-diphosphate (PRPP) to UMP and diphosphate (By similarity)
ENOG4105CZ8	DNA polymerase III is a complex, multichain enzyme responsible for most of the replicative synthesis in bacteria. This DNA polymerase also exhibits 3′ to 5′ exonuclease activity. The beta chain is required for initiation of replication once it is clamped onto DNA, it slides freely (bidirectional and ATP-independent) along duplex DNA (By similarity)
ENOG4105D1X	Specifically dimethylates two adjacent adenosines (A1518 and A1519) in the loop of a conserved hairpin near the 3′-end of 16S rRNA in the 30S particle. May play a critical role in biogenesis of 30S subunits (By similarity)
ENOG4105D27	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates (By similarity)
ENOG4105D5S	Excises uracil residues from the DNA which can arise as a result of misincorporation of dUMP residues by DNA polymerase or due to deamination of cytosine (By similarity)
ENOG4105DA0	Catalyzes the attachment of tyrosine to tRNA(Tyr) in a two-step reaction tyrosine is first activated by ATP to form Tyr-AMP and then transferred to the acceptor end of tRNA(Tyr) (By similarity)
ENOG4105DHW	Required for the insertion and or proper folding and or complex formation of integral membrane proteins into the membrane. Involved in integration of membrane proteins that insert both dependently and independently of the Sec translocase complex, as well as at least some lipoproteins
ENOG4105DJ4	phosphatidylserine decarboxylase
ENOG4105DJV	ribulose-phosphate 3-epimerase
ENOG4105DKZ	GTPase that plays an essential role in the late steps of ribosome biogenesis (By similarity)

Continuation of Table B.3

Eggnog ID	**Annotation**
ENOG4105DRH	Involved in peptide bond synthesis. Stimulates efficient translation and peptide-bond synthesis on native or reconstituted 70S ribosomes in vitro. Probably functions indirectly by altering the affinity of the ribosome for aminoacyl-tRNA, thus increasing their reactivity as acceptors for peptidyl transferase (By similarity)
ENOG4105DTN	riboflavin biosynthesis protein ribF
ENOG4105E5V	Participates in transcription elongation, termination and antitermination (By similarity)
ENOG4105EEE	One of the primary rRNA binding proteins, it binds directly near the 3'-end of the 23S rRNA, where it nucleates assembly of the 50S subunit (By similarity)
ENOG4105EQY	Methylates the class 1 translation termination release factors RF1 PrfA and RF2 PrfB on the glutamine residue of the universally conserved GGQ motif (By similarity)
ENOG4105F8V	Hydrolase, tatD family
ENOG4105G6W	One of the primary rRNA binding proteins, it binds directly to 16S rRNA where it nucleates assembly of the body of the 30S subunit (By similarity)
ENOG4105J80	Produces ATP from ADP in the presence of a proton gradient across the membrane. The gamma chain is believed to be important in regulating ATPase activity and the flow of protons through the CF(0) complex (By similarity)
ENOG4105K46	50S ribosomal protein l27
ENOG4105K4C	This is one of the proteins that binds and probably mediates the attachment of the 5S RNA into the large ribosomal subunit, where it forms part of the central protuberance (By similarity)
ENOG4105K63	Thioredoxin
ENOG4105K77	Forms an intersubunit bridge (bridge B4) with the 23S rRNA of the 50S subunit in the ribosome (By similarity)
ENOG4105K7S	Protein S19 forms a complex with S13 that binds strongly to the 16S ribosomal RNA (By similarity)
ENOG4105K87	One of the primary rRNA binding proteins, it binds specifically to the 5'-end of 16S ribosomal

Continuation of Table B.3

Eggnog ID	**Annotation**
ENOG4105K90	Participates actively in the response to hyperosmotic and heat shock by preventing the aggregation of stress-denatured proteins, in association with DnaK and GrpE. It is the nucleotide exchange factor for DnaK and may function as a thermosensor. Unfolded proteins bind initially to DnaJ
ENOG4105K9U	however, it seems to stimulate more or less all the activities of the other two initiation factors, IF-2 and IF-3 (By similarity)
ENOG4105KAP	The globular domain of the protein is located near the polypeptide exit tunnel on the outside of the subunit, while an extended beta-hairpin is found that lines the wall of the exit tunnel in the center of the 70S ribosome (By similarity)
ENOG4105KAR	One of the proteins that surrounds the polypeptide exit tunnel on the outside of the subunit (By similarity)
ENOG4105KBC	Seems to be the binding site for several of the factors involved in protein synthesis and appears to be essential for accurate translation (By similarity)
ENOG4105KG7	Associates with free 30S ribosomal subunits (but not with 30S subunits that are part of 70S ribosomes or polysomes). Essential for efficient processing of 16S rRNA. May interact with the 5′-terminal helix region of 16S rRNA (By similarity)
ENOG4105KGE	Single strand-specific metallo-endoribonuclease involved in late-stage 70S ribosome quality control and in maturation of the 3′ terminus of the 16S rRNA (By similarity)
ENOG4105KK9	This protein binds to 23S rRNA in the presence of protein L20 (By similarity)
ENOG4105KNM	Produces ATP from ADP in the presence of a proton gradient across the membrane (By similarity)
ENOG4105VH8	Binds as a heterodimer with protein S6 to the central domain of the 16S rRNA, where it helps stabilize the platform of the 30S subunit (By similarity)
ENOG4106U5A	One of the primary rRNA binding proteins, this protein initially binds near the 5′-end of the 23S rRNA. It is important during the early stages of 50S assembly. It makes multiple contacts with different domains of the 23S rRNA in the assembled 50S subunit and ribosome (By similarity)
ENOG4108R5J	This protein binds to the 23S rRNA, and is important in its secondary structure. It is located near the subunit interface in the base of the L7 L12 stalk, and near the tRNA binding site of the peptidyltransferase center (By similarity)

Continuation of Table B.3

Eggnog ID	Annotation
ENOG4108R70	Binds 23S rRNA and is also seen to make contacts with the A and possibly P site tRNAs (By similarity)
ENOG4108RA9	Located at the back of the 30S subunit body where it stabilizes the conformation of the head with respect to the body (By similarity)
ENOG4108UGX	Major role in the synthesis of nucleoside triphosphates other than ATP. The ATP gamma phosphate is transferred to the NDP beta phosphate via a ping-pong mechanism, using a phosphorylated active-site intermediate (By similarity)
ENOG4108UH4	Binds specifically to the SsrA RNA (tmRNA) and is required for stable association of SsrA with ribosomes (By similarity)
ENOG4108UHA	Essential for recycling GMP and indirectly, cGMP (By similarity)
ENOG4108UHH	Located on the platform of the 30S subunit, it bridges several disparate RNA helices of the 16S rRNA. Forms part of the Shine-Dalgarno cleft in the 70S ribosome (By similarity)
ENOG4108UHY	One of the primary rRNA binding proteins, it binds directly to 16S rRNA where it nucleates assembly of the head domain of the 30S subunit. Is located at the subunit interface close to the decoding center, probably blocks exit of the E-site tRNA (By similarity)
ENOG4108UIK	This protein binds directly to 23S ribosomal RNA (By similarity)
ENOG4108UIQ	Methylates the ribose at the nucleotide 34 wobble position in the two leucyl isoacceptors tRNA(Leu)(CmAA) and tRNA(Leu)(cmnm5UmAA). Catalyzes the methyl transfer from S- adenosyl-L-methionine to the 2'-OH of the wobble nucleotide (By similarity)
ENOG4108UJD	30S ribosomal protein S9
ENOG4108UJY	One of the primary rRNA binding proteins, it binds directly to 16S rRNA central domain where it helps coordinate assembly of the platform of the 30S subunit (By similarity)

Continuation of Table B.3

Eggnog ID	**Annotation**
ENOG4108UKE	Interacts with and stabilizes bases of the 16S rRNA that are involved in tRNA selection in the A site and with the mRNA backbone. Located at the interface of the 30S and 50S subunits, it traverses the body of the 30S subunit contacting proteins on the other side and probably holding the rRNA structure together. The combined cluster of proteins S8, S12 and S17 appears to hold together the shoulder and platform of the 30S subunit (By similarity)
ENOG4108UKH	Necessary for efficient RNA polymerase transcription elongation past template-encoded arresting sites. The arresting sites in DNA have the property of trapping a certain fraction of elongating RNA polymerases that pass through, resulting in locked ternary complexes. Cleavage of the nascent transcript by cleavage factors such as GreA or GreB allows the resumption of elongation from the new 3'terminus. GreA releases sequences of 2 to 3 nucleotides (By similarity)
ENOG4108UM5	This protein is one of the early assembly proteins of the 50S ribosomal subunit, although it is not seen to bind rRNA by itself. It is important during the early stages of 50S assembly (By similarity)
ENOG4108UNN	Binds to 23S rRNA. Forms part of two intersubunit bridges in the 70S ribosome (By similarity)
ENOG4108UUX	IF-3 binds to the 30S ribosomal subunit and shifts the equilibrum between 70S ribosomes and their 50S and 30S subunits in favor of the free subunits, thus enhancing the availability of 30S subunits on which protein synthesis initiation begins (By similarity)
ENOG4108UZ0	Binds to the 23S rRNA (By similarity)
ENOG4108VCV	Responsible for the release of ribosomes from messenger RNA at the termination of protein biosynthesis. May increase the efficiency of translation by recycling ribosomes from one round of translation to another (By similarity)
ENOG4108VZM	50s ribosomal protein L10
ENOG4108YY1	This protein is located at the 30S-50S ribosomal subunit interface and may play a role in the structure and function of the aminoacyl-tRNA binding site (By similarity)
ENOG4108YYV	Key enzyme in folate metabolism. Catalyzes an essential reaction for de novo glycine and purine synthesis, and for DNA precursor synthesis (By similarity)

Continuation of Table B.3	
Eggnog ID	**Annotation**
ENOG4108YZX	Binds directly to 23S ribosomal RNA and is necessary for the in vitro assembly process of the 50S ribosomal subunit. It is not involved in the protein synthesizing functions of that subunit (By similarity)
ENOG4108Z04	Located at the top of the head of the 30S subunit, it contacts several helices of the 16S rRNA. In the 70S ribosome it contacts the 23S rRNA (bridge B1a) and protein L5 of the 50S subunit (bridge B1b), connecting the 2 subunits
ENOG4108Z10	Involved in the binding of tRNA to the ribosomes (By similarity)
ENOG4108ZBM	Digests double-stranded RNA. Involved in the processing of primary rRNA transcript to yield the immediate precursors to the large and small rRNAs (23S and 16S). Also processes some mRNAs, and tRNAs when they are encoded in the rRNA operon (By similarity)
ENOG4108ZDX	Binds together with S18 to 16S ribosomal RNA (By similarity)
ENOG4108ZEF	Reversibly transfers an adenylyl group from ATP to 4'- phosphopantetheine, yielding dephospho-CoA (dPCoA) and pyrophosphate (By similarity)
ENOG4108ZMD	Phosphorylation of dTMP to form dTDP in both de novo and salvage pathways of dTTP synthesis (By similarity)
ENOG4108ZQD	Catalyzes the phosphorylation of the 3'-hydroxyl group of dephosphocoenzyme A to form coenzyme A (By similarity)
ENOG4108ZT0	50S ribosomal protein l17
End of Table B.3	

TABLE B.4: **The number of sequence pairs for each gene in the 50 and 201 pairs subsets:** This table illustrates genes that are found for multiple PIDs. The highest number of pairs for the 201 pairs subset is from ENOG4105BZH which is annotated as the gene that provides the precursors necessary for DNA synthesis by catalyzing the biosynthesis of deoxyribonucleotides from the corresponding ribonucleotides. The highest number sequence pairs for the 50 pairs subset is from RF00177 which is 16s rRNA (aka small subunit ribosomal RNA).

Start of Table B.4		
Gene	**201 Pairs**	**50 Pairs**
SeC	1	0
RF01854	3	0
ENOG4105D5S	13	9
ENOG4105EQY	19	7

Continuation of Table B.4

Gene	201 Pairs	50 Pairs
ENOG4105C20	20	10
RF00001	22	167
ENOG4105KAP	26	18
ENOG4105C5S	27	8
ENOG4105CRK	28	4
ENOG4105CV1	29	15
ENOG4105C38	30	2
ENOG4108UGX	32	16
ENOG4108UKH	35	12
ENOG4105C9M	35	13
ENOG4105C6P	35	17
ENOG4105CE9	36	11
ENOG4108ZEF	37	14
ENOG4105C8K	37	18
ENOG4105C24	40	16
ENOG4105C8N	40	16
ENOG4105D27	40	17
ENOG4105C8T	40	19
ENOG4105C3G	40	22
ENOG4108UUX	41	13
ENOG4105CA4	41	16
ENOG4105CPH	41	27
ENOG4105CGJ	42	11
ENOG4105BZ6	42	14
ENOG4105CIM	42	16
ENOG4105DJV	42	16
ENOG4105C22	43	8
ENOG4108YY1	43	22
ENOG4105C7D	43	24
ENOG4105C0V	44	11
ENOG4105DRH	44	14
ENOG4108UH4	44	16
ENOG4105CPM	44	18
ENOG4105CJS	45	9
ENOG4108UHA	45	22
ENOG4105CDU	46	13
ENOG4105E5V	46	16
ENOG4105C2M	47	17
ENOG4105K77	47	19

Continuation of Table B.4

Gene	**201 Pairs**	**50 Pairs**
ENOG4108YZX	48	30
ENOG4105CA1	49	14
ENOG4108UZ0	49	18
ENOG4105KK9	49	21
ENOG4108VCV	50	8
ENOG4105CA9	50	14
ENOG4105C77	50	16
ENOG4105EEE	50	18
ENOG4105CZ5	51	11
ENOG4105BZ5	51	13
ENOG4105D1X	51	14
ENOG4105C62	51	19
ENOG4105C65	51	21
ENOG4108UM5	51	22
ENOG4105DKZ	51	23
ENOG4108UJY	51	30
ENOG4105K46	51	34
ENOG4105CI6	52	13
ENOG4105CCJ	54	11
ENOG4105C9R	54	15
ENOG4105C64	54	19
ENOG4105K87	55	26
ENOG4108R5J	55	29
ENOG4105CSM	56	12
ENOG4108RA9	56	23
ENOG4105CGR	57	18
ENOG4108UHH	57	27
ENOG4105G6W	57	35
ENOG4105K4C	58	15
ENOG4105CSS	58	18
ENOG4108UJD	58	21
ENOG4105CDK	59	12
ENOG4105C8D	59	13
ENOG4105CB9	59	15
ENOG4105C4S	59	27
ENOG4105CFD	59	39
ENOG4105C5T	61	17
ENOG4105CKE	61	18
ENOG4105KAR	61	25

Continuation of Table B.4

Gene	201 Pairs	50 Pairs
ENOG4108UHY	61	33
ENOG4105C31	63	12
ENOG4108ZDX	64	20
ENOG4105K9U	64	40
ENOG4105C3H	65	14
ENOG4105C17	66	14
ENOG4105J80	66	20
ENOG4108Z04	67	28
ENOG4108UIK	68	35
ENOG4105C70	69	20
ENOG4105CTF	69	26
ENOG4105C3M	70	17
ENOG4105CDG	70	36
ENOG4105CU0	71	11
RF00023	72	10
ENOG4105CU7	72	21
ENOG4108R70	72	47
ENOG4105CFG	73	35
ENOG4106U5A	74	30
ENOG4105CEJ	74	35
ENOG4105F8V	75	24
ENOG4105CW6	75	38
ENOG4105CJ9	77	44
ENOG4108Z10	81	38
ENOG4105K7S	81	49
ENOG4105DTN	85	23
ENOG4105VH8	92	39
ENOG4108UNN	93	57
ENOG4108ZT0	94	32
ENOG4108VZM	98	26
ENOG4105C6A	101	22
RF00169	104	98
ENOG4105CGV	105	87
ENOG4108UKE	107	53
ENOG4105CHV	124	33
ENOG4108ZBM	136	24
ENOG4105CZ8	138	25
RF02541	155	190
ENOG4105KGE	156	37

Continuation of Table B.4

Gene	**201 Pairs**	**50 Pairs**
Ser	186	0
ENOG4105K90	187	45
Thr	188	0
ENOG4105KG7	197	49
Arg	226	0
Met	232	0
Leu	234	0
ENOG4105DA0	234	66
His	242	0
Gln	245	0
Tyr	249	0
Asn	250	0
Lys	255	0
Cys	267	0
RF00177	273	**257**
Val	274	0
Glu	275	0
ENOG4105C9G	278	72
ENOG4105KNM	287	72
Trp	300	0
Pro	352	0
ENOG4105C4J	368	109
Ile	382	0
Ala	393	0
ENOG4105DHW	404	104
Phe	421	0
ENOG4105CWT	440	113
ENOG4105CEY	458	143
ENOG4105C1H	460	121
Asp	484	0
Gly	524	0
ENOG4105C90	765	200
ENOG4105BZH	**828**	212

End of Table B.4

TABLE B.5: **The number of pairs of mRNA, ncRNA and tRNA for each PID subset:** The 50 pairs subset only has mRNA and ncRNA sequences whereas the 201 pairs subset includes tRNA sequences. This shows the over representation (greater then 20% of the pairs) of tRNA sequence pairs for PID values higher then 61% and the lack of ncRNAs for low PIDs.

Start of Table B.5

PID	**mRNA 50 Pairs**	**ncRNA 50 Pairs**	**mRNA 201 Pairs**	**ncRNA 201 Pairs**	**tRNA 201 Pairs**
1	10	0	40	0	0
2	10	0	40	0	0
3	10	0	40	0	0
4	10	0	40	0	0
5	10	0	40	0	0
6	10	0	40	0	0
7	10	0	40	0	0
8	10	0	40	0	0
9	10	0	36	0	0
10	17	0	23	0	0
11	28	0	41	0	0
12	41	0	135	0	0
13	50	0	146	0	0
14	50	0	168	0	0
15	50	0	201	0	0
16	50	0	201	0	0
17	50	0	201	0	0
18	50	0	201	0	0
19	50	0	201	0	0
20	50	0	201	0	0
21	50	0	201	0	0
22	50	0	201	0	0
23	50	0	201	0	0
24	50	0	201	0	0
25	50	0	201	0	0
26	50	0	201	0	0
27	50	0	200	1	0
28	49	1	198	3	0
29	50	0	198	3	0
30	49	1	198	3	0
31	49	1	195	6	0
32	49	1	190	10	32
33	50	0	194	7	0

Continuation of Table B.5

PID	mRNA 50 Pairs	ncRNA 50 Pairs	mRNA 201 Pairs	ncRNA 201 Pairs	tRNA 201 Pairs
34	49	1	196	5	0
35	46	4	195	6	0
36	50	0	189	12	0
37	49	1	190	11	0
38	49	1	190	11	0
39	50	0	197	4	0
40	49	1	199	2	0
41	49	1	196	5	0
42	48	2	194	6	42
43	49	1	199	1	43
44	50	0	194	7	0
45	46	4	197	4	0
46	49	1	198	3	0
47	49	1	195	6	0
48	49	1	196	4	1
49	50	0	198	2	1
50	50	0	200	0	1
51	49	1	199	1	1
52	49	1	196	3	2
53	47	3	197	0	4
54	50	0	199	0	2
55	49	1	195	2	4
56	47	3	196	1	4
57	48	2	195	3	3
58	49	1	181	0	20
59	49	1	187	3	11
60	48	2	179	0	22
61	47	3	164	3	34
62	48	2	157	2	42
63	47	3	153	3	45
64	47	3	123	2	76
65	46	4	128	8	65
66	42	8	111	6	84
67	47	3	111	10	80
68	42	8	64	10	127
69	35	15	83	13	105
70	34	16	61	22	118
71	35	15	45	19	137

Continuation of Table B.5

PID	**mRNA 50 Pairs**	**ncRNA 50 Pairs**	**mRNA 201 Pairs**	**ncRNA 201 Pairs**	**tRNA 201 Pairs**
72	32	18	31	32	138
73	29	21	44	23	134
74	31	19	40	19	142
75	29	21	30	20	151
76	32	18	28	31	142
77	33	17	22	28	151
78	33	17	22	18	161
79	34	16	12	19	170
80	30	20	15	21	165
81	26	24	12	20	169
82	26	24	14	10	177
83	26	24	6	10	185
84	22	28	6	6	189
85	18	32	18	14	169
86	24	26	10	7	184
87	23	27	2	14	185
88	28	22	4	9	188
89	23	27	3	15	183
90	26	24	10	24	167
91	18	32	9	5	187
92	31	19	5	9	187
93	25	25	6	11	184
94	22	28	13	19	169
95	23	27	5	4	192
96	25	25	1	3	197
97	25	25	1	4	196
98	31	19	70	13	118
99	36	14	3	1	197
100	42	8	3	2	196

End of Table B.5

ORF Number	ENOG4105C1H	ENOG4105CWT	ENOG4105DKZ
orf178332101[A]	✓	✓	
orf180067068[A]	✓	✓	
orf179254753	✓		✓
orf179642852	✓	✓	
orf92089854	✓		✓
orf95036926	✓		✓

ORF Number	Phylum	Genus and Species
orf178332101[A]	NA*	*Thermobaculum terrenum*
orf180067068[A]	Choloroflexi	*Thermorudis peleae*
orf179254753	Choloroflexi	*Thermoflexus hugenholtzii*
orf179642852	Choloroflexi	*Thermomicrobium roseum*
orf92089854	Choloroflexi	*Leptolinea tardivitalis*
orf95036926	Choloroflexi	*Longilinea arvoryzae*

Eggnog HMM	Annotation
ENOG4105C1H	Exhibits a very high intrinsic GTPase hydrolysis rate. Involved in the addition of a carboxymethylaminomethyl (cmnm) group at the wobble position (U34) of certain tRNAs, forming tRNA- cmnm(5)s(2)U34 (By similarity)
ENOG4105CWT	An essential GTPase that binds both GDP and GTP, with rapid nucleotide exchange. Plays a role in 16S rRNA processing and 30S ribosomal subunit biogenesis and possibly also in cell cycle regulation and energy metabolism (By similarity)
ENOG4105DKZ	GTPase that plays an essential role in the late steps of ribosome biogenesis (By similarity)

TABLE B.9: **ORFs with duplicate annotations: *Top Table***: Shows the sequences (identified by orf number) that are top scoring sequence for a genome for two different genes HMMs. ***Middle Table***: Shows the taxonomic information for each of the sequence from the top table. ***Bottom Table***: Shows the EggNOG annotation for each HMM that has sequences that are in also the top scoring match for another HMM.

A=in both 50 and 200 PID subset, *=disputed phylogeny hence the NA in phylum but a literature review currently indicates it is likely to be from the Chloroflexi phylum. (111)

Gene	PID from reference alignment	PID from pairwise alignment
ENOG4105C4J	60%	65%
ENOG4105KG7	42%	59%
RF00177	80%	83%
RF00169	40%	53%

TABLE B.10: **Comparison of PIDs of the same sequence pairs using different alignment strategies:** Nucleotide PID calculations of sequence pairs from each gene showing the difference in PID from the reference to the pairwise alignment. All PID values from the pairwise, sequence similarity amximised, alignment are higher then the PID values from the reference alignment. The reference alignment is made by aligning all sequences for a gene to the corresponding HMM or CM. The pairwise alignment is made by aligning pairs of sequences to each other with ggsearch.

	G0-G0	**G0-G1**	**G0-G2**	**G1-G1**	**G1-G2**	**G2-G2**
L0-L0	0	0	0	0	0	**1**
L0-L1	0	0	0	0	0	0
L1-L1	1	7	8	5	10	0
L1-L2	0	0	1	0	0	0
L2-L2	0	1	4	3	8	1

TABLE B.11: **The combinations of G+C and length for 40% PID:** The G+C content and sequence length groups each sequence pair belongs to for pairs in the 50 pairs subset that have 40% PID. L0: Short sequence, L1: Medium length sequence, L2: Long length sequence. G0: Low G+C content, G1: Mid G+C content, G2: High G+C content. The bold number shows the long G0-G2 short length sequence pair referred to in the discussion.

Appendix C

Supplementary Github links to files and code

https://github.com/Gardner-BinfLab/twilight_zone.git	/pipeline

TABLE C.1: **GitHub code and pipeline instructions:** All custom code (Perl, Bash or R) and a description of the computational pipeline (README.md) is available at this GitHub link in the pipeline directory.

BLASTn v2.6.0	ggsearch36/ssearch36/ssearch34	nhmmer v3.1b2
-gapopen 10	-f 10	–toponly
-gapextend 6	-g 6	–dna
-penalty -4	-r "+5/-4"	–F1 0.3
-reward 5	-d 0	–F2 0.1
-strand plus	-3	–F3 0.03
-task blastn	-m 3	
	-n	

TABLE C.2: **Alignment parameters:** Parameters and their corresponding values for each of the alignment algorithms used.

Link A:	Genomes and Taxonomy Table	linkA.tsv
Link B:	Eggnog HMMs Gene Annotations	linkB.tsv
Link C:	Sequence alignment to ENOG4105BZ5 HMM of two duplicates and example real homologs	dup_align.afa
Link D:	Taxonomic Tree	taxonomic_tree.pdf
Link E:	Duplicate Sequence example of low complexity	low_complexity.fasta
Link F:	Randomly generated sequence	rand_seq.txt

TABLE C.3: **GitHub links to large tables and supplementary files:** All files can be found in the main directory https://github.com/Gardner-BinfLab/twilight_zone.git