FUNCTION PREDICTION IN RNA AND PROTEINS

A Dissertation Submitted to the Faculty of

The Graduate School of Biomedical Sciences
Baylor College of Medicine

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

by

ILYA B. NOVIKOV

Houston, Texas

June 3, 2019

APPROVED BY THE DISSERTATION COMMITTEE

Olivier Lichtarge, M.D., Ph.D.
Chair

Tom Cooper, Ph.D.

Edward Nikonowicz, Ph.D.

Timothy Palzkill, Ph.D.

B. V. Venkatar Prasad, Ph.D.

Lynn Zechiedrich, Ph.D.

APPROVED BY THE BIOCHEMISTRY AND MOLECULAR BIOLOGY GRADUATE PROGRAM

B. V. Venkatar Prasad, Ph.D.
Director of Graduate Studies

APPROVED BY INTERIM DEAN OF THE GRADUATE SCHOOL OF BIOMEDICAL SCIENCES

Adam Kuspa, Ph.D.

Date

Acknowledgments

The author gratefully acknowledges Angela Wilkins for key contributions to this work, both in terms of experimental design and conceptualization. Additionally, the author wishes to thank Ben Bachman, Rhonald Lua, David Murciano, Panos Katsonis, and Brigitta Wastuwidyaningtyas for their expertise and helpful feedback.

Abstract

In order to develop novel disease therapies, one must first understand how individual proteins and nucleic acids function within the cell. In practice, function determination consists of two related, but distinct goals. First, given a protein (or nucleic acid) we seek to discover its function relative to the other molecules in the cell. These functions include enzymatic activity, interactions with other macromolecules, structural role, and so on. Second, having discovered the molecule's functional role, we next seek to identify individual residues (or nucleotides) that contribute to it, such as residues that make up catalytic sites and binding interfaces. In this manner, comprehensive functional annotation is a two-step process: we must both determine a molecule's role within the cell, and the role of individual residues within the molecule. In this work, we explore the problem of function annotation from both of these perspectives.

First, we address the problem of functional discrimination in non-coding (nc)RNAs on single-nucleotide level. Functional ncRNA are nucleotide sequences of varied lengths, structures, and mechanisms that ubiquitously influence gene expression and translation, genome stability and dynamics, and human health and disease. Here, to shed light on their functional determinants, we seek to exploit the evolu-

tionary record of variation and divergence read from sequence comparisons. The approach follows the phylogenetic Evolutionary Trace (ET) paradigm, first developed and extensively validated on proteins. Here, we assigned a relative rank of importance to every base in a study of 1070 functional RNAs, including the ribosome, and observed evolutionary patterns strikingly similar to those seen in proteins, namely, (1) the top-ranked bases clustered in the structure. (2) In turn, these clusters mapped functional regions for catalysis, binding proteins and drugs, post-transcriptional modification, and deleterious mutations. (3) Moreover, the quantitative quality of these clusters correlated with the identification of functional regions. (4) As a result of this correlation, smoother structural distributions of evolutionary important nucleotides improved functional site predictions. Thus, in practice, phylogenetic analysis can broadly identify functional determinants in RNA sequences and functional sites in RNA structures, and reveal details on the basis of RNA molecular functions. As example of application, we report several previously undocumented and potentially functional ET nucleotide clusters in the ribosome. This work is broadly relevant to studies of structure-function in ribonucleic acids. Unlike other current methods, ET does not impose size limits on input alignments, and by quantifying divergence with respect to evolutionary history it is more accurate for functional site discrimination than information entropy. This allows ET to be applied, in a practical manner, to full-length alignments of large molecules such as HIV and the ribosome.

The next problem addressed in this work is that of protein function annotation in the context of the proteome We present two related approaches for automatically

annotating proteins using the popular classification database, the Gene Ontology (GO), as the sole source of input features. Gene Ontology provides functional annotations for proteins by assigning them labels, referred to as GO terms. GO terms describe a proteins function, role, and location within the cell. We sought to use existing Gene Ontology annotations as the principal source of input data for building predictive models that infer future annotations. In the first approach, we extract the term annotations as a term-protein matrix, and use Non-negative Matrix Factorization (NMF) to suggest novel term-protein annotations. We show that this representation is stable in ten-fold cross-validation, and that it has predictive power in time-stamped trials. In the second approach, we use Resnik semantic similarity to build a pairwise distance matrix for annotated proteins, and then convert the distance matrix into a protein-protein network. To reason on this representation, we label protein nodes with known functions, and then use graph diffusion to propagate the labels to the unlabeled nodes. We show that this graph representation is also stable and self-consistent, and then use time-stamp validation to show that it, too, has predictive power. Together these data show that GO can be readily distilled into a form amenable to automated reasoning, and that it has predictive power.

Contents

List of Figures

List of Tables

Chapter 1

Overview of Functional Non-coding RNA Families

One of the greatest scientific achievements of the 20th century is the discovery of the Central Dogma of molecular biology [1, 2]. As outlined for the first time by Francis Crick, the Central Dogma describes the relationship between 3 major polymer types found in all living systems: deribonucleic acids (DNAs) store genetic information, proteins perform the assorted variety of cellular functions, and ribonucleic acids (RNAs) serve the short-lived intermediate between the two. In this canonical view, RNA was regarded as too fragile to serve as a long-term information repository (due to the reactive 2' hydroxyl), and too invariant (only 4 bases) to perform protein-like functions. Thus, traditional role of RNAs was as a messenger (m)RNA that connects DNA to protein.

However, the notion that RNA can play other protein-like roles appeared almost immediately after the establishment of the Central Dogma [3]. This idea gained significant traction in the 1980s when self-splicing introns were discovered in *Tetrahymena* rRNA [4], and bacterial RNAse P was confirmed to have a RNA catalytic core [5]. These molecules featured catalytically-active RNAs that did not encode protein but instead functioned like one. These discoveries paved way for further inquiry into functional non-coding (fnc)RNAs, and led to rapid expansion in the number of known functional RNAs. Today, fncRNAs include several classes such as ribozymes, riboswitches, transfer (t)RNA and ribosomal (r)RNA, microRNAs, small nucleolar (sno)RNAs, small nuclear (sn)RNAs, and the long non-coding (lnc)RNAs.

In more detail, functional RNAs can be categorized into several broad groups. First are the classic RNAs involved in the core aspects of translation. Principally,

this group includes ribosomal and transfer RNAs [6]. Ribosome is approximately $\frac{2}{3}$ RNA by weight, and RNA serves both as the structural foundation of the ribosome and its catalytic core. Working in tandem with rRNA are the much smaller transfer (t)RNAs, which deliver aminoacids to the ribosome for incorporation into the growing peptide chain. Another RNA machine associated with translation is RNAse P, an enzyme that promotes maturation of tRNAs by cleaving the 5' end of the precursor tRNAs [5, 7]. Likewise, two other families, small nucleolar (sno)RNAs and small nuclear (sn)RNAs, are also closely linked to translation and transcription. SnoRNAs guide post-transcriptional modification of nascent ribosomal RNA (2-O-ribose methylation and pseudouridylation), which is necessary for rRNA maturation [8]. Meanwhile, snRNAs are best known for being a constituent part of the spliceosome, where they appear to guide splice site recognition [9].

Another group of classic fncRNAs are the regulatory motifs embedded in untranslated regions of mRNA molecules, most notably riboswitches. Riboswitches are often found in bacterial genes involved in metabolite production [10]. One representative example is the glmS riboswitch [11], located in the 5' UTR of the *glmS* gene that encodes the enzyme responsible for production of glucosamine-6-phosphate (GlcN6P). When concentration of GlcN6P is high, it binds to the the glmS riboswitch inducing it to cleave the *glmS* transcript. In this manner, the riboswitch regulates expression of *glmS*. Other similar regulators include ribozymes [12], such as the hammerhead ribozyme that participates in viral rolling circle replication, and self-splicing introns [13].

Discovered after the classic RNAs were the novel calsses of small regulatory

RNAs and long non-coding (lnc)RNAs. Small regulatory RNAs include microRNAs (miRNAs) and small interfering RNAs (siRNA), which regulate gene expression by transcript suppression [8]. In their mature form, mi/siRNAs are part of the RNA-induced silencing complex (RISC), mediating RISC binding to mRNA targets that are complementary to the mi/siRNA sequence. Once bound, RISC cleves the mRNA target. There are hundreds of known mi/siRNAs, and they are perhaps best known for their role in development. Finally, the most nascent class of fRNAs are long non-coding RNAs [14]. While thousands of putative lncRNAs exist, few have been functionally characterized (most notably *Xist*, which controls inactivation of the X chromosome in mammals), and there is much debate about their mode of function.

Thus, functional non-coding RNAs are diverse, varied, and involved in multiple core functions of the cell. Furthermore, a number of these molecules have already been associated with human disease [15–18]. It is important to better understand and modulate these molecules, and to do so we must first understand what determines their function on single-nucleotide level. The traditional experimental methods to solve this problem, such as scanning mutagenesis, scale poorly with the growing number of new fncRNAs discovered every year (Fig 1.1). Meanwhile, traditional computational methods are dominated by the problem of structure prediction, not functional delineation.

In Chapter 2, we address the problem of predicting functional determinants in these functional non-coding RNAs using Evolutionary Trace (ET), a sequence analysis tool that scans the evolutionary history of a given RNA to identify nu-

cleotides that are likely critical to function. In Chapter 2, we provide a theoretical basis for ET's novel use in fncRNA, examine in detail its application to two fRNA model systems (the hammerhead ribozyme and the ribosome), and show that the method is general across diverse families of fRNAs. We also explore a number of optimization techniques for improving ET's prediction accuracy in RNA, and suggest four novel functional sites in the bacterial ribosome.

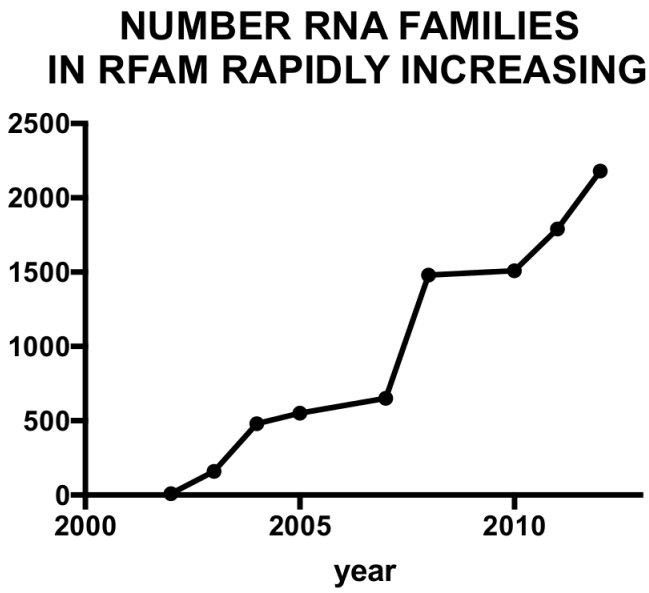

Figure 1.1: Rfam database, the principle repository for annotated fn-cRNA families, has been growing rapidly [19].

Chapter 2

Evolutionary Trace Method Defines Functional Sites Common to RNA Families

This work is under review as Ilya Novikov, Angela Wilkins, and Olivier Lichtarge. (2019) Evolutionary Trace method defines functional sites common to RNA families, *PLOSCB*.

Author contribution: Conceived and designed experiments: IN, AW, OL; Performed the experiments: IN; Analyzed the data: IN, AW. Wrote the paper: IN, AW, OL.

2.1 Introduction

Functional non-coding (fnc)RNAs are a broad class of functional macromolecules that regulate transcription and translation, maintain genome stability [20], and play a role in diseases. They are found across evolution and include both classical as well as several new forms discovered over the past 30 years. The well-known classical RNAs primarily concern translation: they are ribosomal (r)RNA, transfer (t)RNA, small nucleolar (sno)RNA, and the tRNA maturation enzyme RNAse P. The novel RNA classes span self-splicing ribozymes that control viral replication, riboswitches that regulate small molecule metabolism in bacteria, small regulatory RNAs (microRNAs) that regulate mRNA translation in eukaryotes, and, most recently, long non-coding (lnc)RNAs that impact pre- and post-transcriptional gene expression [14]. Thus, functional non-coding RNAs are diverse and contribute significantly to cell metabolism. Critically, fncRNAs have been linked to human disease. For example, mutations in mitochondrial RNAse P are associated with cartilage-hair hypoplasia [21], deletion of promoter that drives expression of HBII-85 snoRNAs contributes to Prader-Willi syndrome [22], and mutations in hTR, RNA component of DNA telomerase, promote Dyskeratosis congenita [16]. Furthermore, small regulatory RNA are perturbed in cancer, in cardiovascular diseases, and neurodegenerative disorders [23], and studies have shown that fncRNA expression is significantly disturbed in cancer cell lines [17]. Long non-coding RNA MALAT1 has been directly linked to metastasis in lung and gastric cancer [18]. These and other fncRNAs represent an entirely new class of druggable targets. Indeed, a number of inhibitors have already been developed to target pathogenic

fncRNA, including riboswitches [24, 25] and the ribosome [26]. Given the growing recognition of the role of fncRNA in human health [27], it is important to understand the determinants of function in these molecules.

To understand fncRNA structure and function and target them for therapy, a central question is which nucleotides in a given molecule contribute to function? Answers have thus far relied on structure determination and targeted mutagenesis. First, secondary or tertiary RNA structures are solved by any number of wet-lab techniques, such as x-ray crystallography, NMR, and enzymatic or chemical probing [28], or via in silico algorithms [29–33]. Based on the structure model, specific nucleotides may then be targeted for mutagenesis, as in [34]. This classical experimental paradigm is resource intensive and contingent on suitable biochemical assays, cell lines, and viable mutants.

In protein research, the similar challenge of identifying functionally important amino acid residue had been effectively addressed by the predictive computational methods, most notably Evolutionary Trace, which is the single most-validated approach [35]. However, in RNA research, there are currently no well-validated computational alternatives to the experimental paradigm (one notable exception is the protein-centric ConSurf web-tool that recently added the ability to score conservation of nucleic acid sequences [36], but did not provide thorough validation for RNA). Because the field is so nascent, most RNA sequence analysis tools, such as GERP++ and PhastCons [37, 38], are used primarily in genomic context to identify novel exons or ncRNAs, and in practice, they are not applicable to single-nucleotide functional analysis of individual RNA molecules.

Furthermore, the traditional purpose of sequences analysis in RNA has been to model secondary and tertiary structure via detection of canonical Watson-Crick base pairing. The first studies of homologue co-variation led to secondary structures of tRNA [39], 5S rRNA [40], and self-catalytic introns [41]. Structure prediction with aid of RNA sequence analysis further evolved with context-free grammar algorithms [27], and the recent advances in the field deal with prediction of non-canonical long-range tertiary contacts in larger molecules [42]. Unlike ET, these methods are primarily aimed at structure prediction, and do not directly provide analysis of evolutionary importance on single-nucleotide level.

To address this need, we sought to adapt Evolutionary Trace [43, 44] to predict functional nucleotides in RNA from their evolutionary history. Evolutionary Trace is a method to identify functional important residues in proteins. It correlates sequence variations with evolutionary divergences in order to rank sequence positions as more (or less) important to function (2.1). In so doing, ET makes two assumptions. First, that sequence variations during evolution and speciation are akin to sampling the sequence-function space via wet lab mutations. Second, that the depth of divergence between two sequences is commensurate with their functional difference, that is, this depth is a quantitative assay of functional distance. If so, a systematic tally of the variations, at any given position of a multiple sequence alignment, that track mostly with deep (or small) phylogenetic divergence, enable ET to assign a greater (or lesser) relative rank of evolutionary importance to each sequence position. More recently, it was recognized that such systematic coupling of variations in sequence space (genotype) with variations in

evolution (fitness space) can be formally recast as a gradient of the evolutionary mapping of genotypes onto the fitness landscape [45]. Viewing ET as the gradient of the evolutionary landscape, presumably a foundational feature of biology, helps explain that the relatively simple in silico process of tracing phylogenetic trees and alignments of homologous sequences (see Methods) leads to varied and useful insights into the molecular basis of protein function. By targeting mutations to top-ranked sequence positions (so-called ET residues), ET-guided studies identify protein-protein interaction interfaces [46–48], allosteric [49] and ligand binding sites [50], recode ligand specificity [49], designed functionally-active peptides [51], and on a structural proteomic scale computationally predict the function of orphan proteins [52, 53].

A natural question is whether similar insights might be gathered for RNA by translating the ET formalism from amino acids sequences to nucleotide sequences. This is readily testable since in proteins top-ranked ET residues have well established general properties that underpin the methods successes: (1) ET can rank amino acid sequence positions from most to least important, such that those in top 30th percentile are called ET residues. (2) These ET residues cluster in the three-dimensional structure of the molecule [54] and (3) overlap its functional sites [46]. (4) Critically, the quality of the structural clustering of ET residues informs with quality of functional site overlap [55]. And finally, (5) the quality of overlap can be improved via optimized sequence selection that maximizes ET clustering [55] and minimizes rank differences of neighboring residues (a structural smoothing of ET ranks) [56].

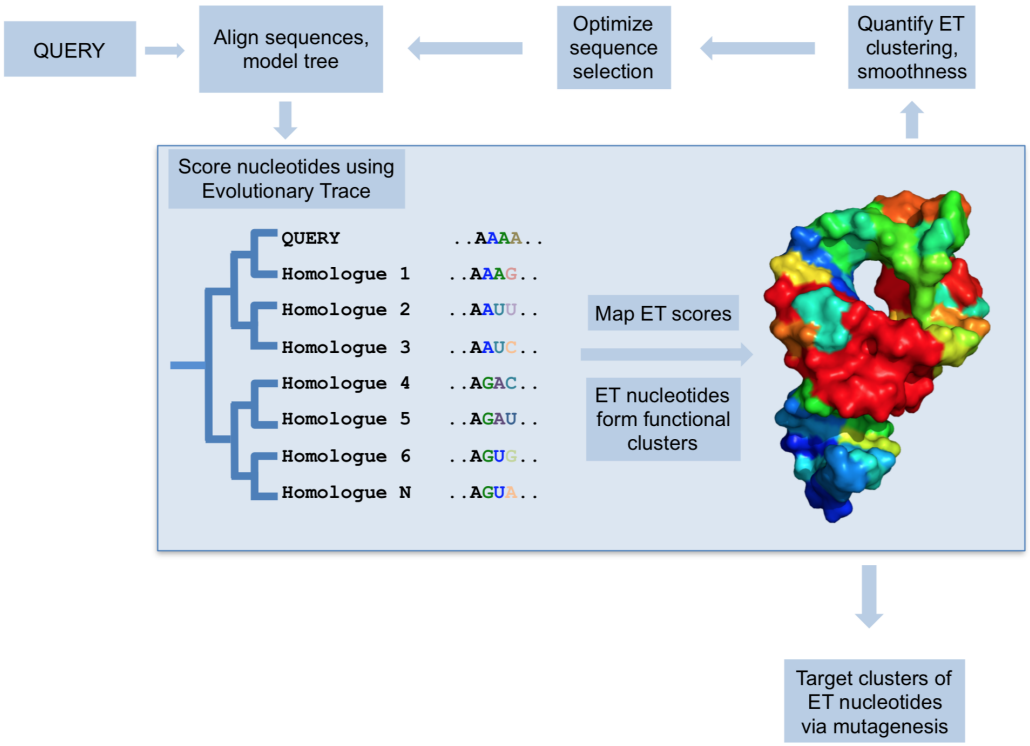

Figure 2.1: The Evolutionary Trace model. For a set of homologues, ET quantifies correlation between phylogenetic tree divergence and sequence variation. ET nucleotides, where the correlation is highest, are considered evolutionarily- and functionally-important. They cluster on the structure and predict functional sites. Furthermore, the quality of structural clustering by ET nucleotides can be measured, and then optimized to improve functional site prediction.

Therefore, to generalize the use of ET to RNA sequences, we sought to test whether ET nucleotides exhibit these five properties. We applied our tests to a representative set of RNA molecules from the Rfam database [19] (Fig 2.2A), and found that ET bases obey the same general rules as ET residues. In particular, we focused on a subset of well-characterized RNAs with known tertiary structures (Fig 2.2B), which account for 7% of our test set and are also fairly representative of overall RNA biology. In practice, the data show that Evolutionary Trace can be readily applied to multiple sequence alignments of homologous fncRNAs to identify nucleotides of functional importance.

2.2 Evolutionary Trace Theory and Methods

2.2.1 Measuring Nucleotide Importance with Real-value ET

To measure nucleotide importance, we use Evolutionary Trace (2.1). The first step in the ET analysis is to construct a representative multiple sequence alignment (MSA) for the query sequence and its homologues. Here, we use the manually-curated seed alignments from the Rfam database [19], that each have at least 10 canonical sequences. The alignment is used to construct a UPGMA phylogenetic tree, and the two are then traced. ET iterates through the sequence columns and assigns a rank based on how closely the sequence variation within the column correlated with tree branching. The first-generation ET algorithm [43, 46] expresses the rank as:

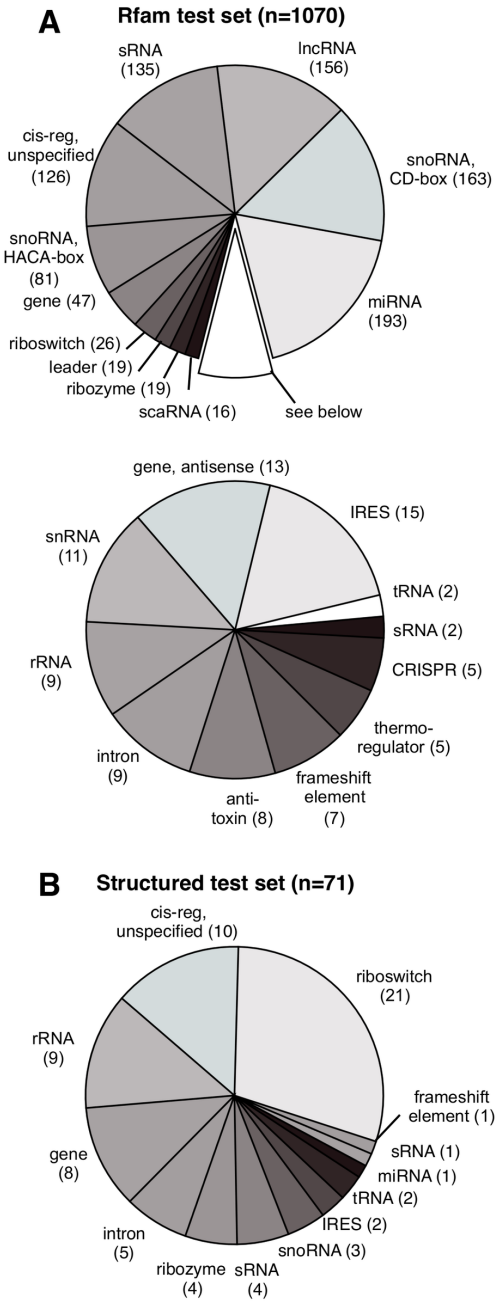

Figure 2.2: Rfam test set represents a broad selection of functional RNAs. Shown in (A) are Rfam families we used in our test set. In (B) is a subset of Rfam test families that map to high-resolution structures in the Protein Data Bank (PDB), allowing us to study three-dimensional clustering by ET nucleotides.

$$r_i = 1 + \sum_{n=1}^{N-1} \delta(n) \begin{cases} \delta(n) = 0 & \text{if node } n \text{ is invariant in every group } g \\ \delta(n) = 1 & \text{otherwise} \end{cases} \quad (1)$$

where r_i is the rank of the residue at position i, N is the total number of sequences in the tree, and $N - 1$ is the number of nodes. To compute r_i, we iterate through every node n staring with one closest to the root ($n = 1$), and divide the sequences in the alignment into subgroups g based on the topology of the tree. Because the tree is binary, at node level n, the tree is divided into $g = n$ groups. We assign 0 to $\delta(n)$ if the residue at position i is invariant within all sequence groups g, and 1 otherwise. Thus, evolutionarily important nucleotides that are fixed within major branches will receive a lower absolute rank r_i than residues that continue to vary as the tree is traversed.

This approach produces integer ranks, and suffers from treating each node as equally important, which is not always true. To address this, we developed real-value ET (rvET) [35,44], an extension of the basic method, that uses information entropy to weight phylogenetic branches according to their sequence conservation:

$$r_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \left\{ - \sum_{a=1}^{20} f_{ia}^{g} ln f_{ia}^{g} \right\} \quad (2)$$

where f_{ia}^{g} is frequency of an amino (or nucleic) acid a found within the sequence group g. Now, as we traverse the tree, we sum up and then average the information entropy for each of the sub-alignments g observed when we split the tree into n

nodes. This allows us to produce better resolved ranks r_i that are more resistant to sequence inconsistencies, while still taking into account the phylogenetic history of the tree. In this manner, we arrive at a relative ranking of evolutionary importance for every position in the alignment. In practice, we normalize ranks into percentile ranks, or coverage. The coverage of 5% includes the top 5% of the most highly ranked nucleotides, and so on. Here, we refer to ET nucleotides as nucleotides within 35% ET rank coverage.

Measuring Nucleotide Clustering and Overlap

To measure structural clustering by ET nucleotides, we developed the concept of Selection Clustering Weight (SCW), described in detail in [57]. Briefly, for a set of nucleotides S, Selection Clustering Weight, w, is the number of structural contacts formed by the members of S. To calculate w, we present the structure in form of adjacency matrix A:

$$A(i,j) = \begin{cases} 1 & \text{if } d(i,j) < 4\text{Å} \\ 0 & \text{if } d(i,j) > 4\text{Å} \end{cases} \tag{3}$$

where d is the distance between any two nucleotides i and j, and a contact is denoted by $A(i,j) = 1$ if d is 4Å or less. Using selection function $S(x)$ (which returns 0 if nucleotide x is not found in selection, and 1 otherwise), we iterate over A and calculate w:

$$\omega = \sum_{i>j}^{L} S(i)S(j)A(i,j) \tag{4}$$

To assess the statistical significance of ω, we compare it to the mean expected

clustering weight, $\langle\omega\rangle$, by a random set of nucleotides of the same size as S. We express the difference between clustering weight of nucleotide set S, and random nucleotides, in form of a clustering z-score z_c:

$$z_c = \frac{\omega - \langle\omega\rangle}{\sigma} \tag{5}$$

where σ is the standard deviation of $\langle\omega\rangle$. Both $\langle\omega\rangle$ and σ can be calculated analytically, as explained in [57]. Using this procedure, we calculated the statistical significance of clustering by ET nucleotides.

Similar to clustering z-score, we introduce overlap z-score z_o to assess how well ET nucleotides predict functionally-relevant sites. Given a pre-defined set of functional nucleotides of size M and a set of ET nucleotides of size n in a molecule of length N, we can use hypergeometric distribution to calculate mean expected overlap between the two:

$$m = n\frac{M}{N} \tag{6}$$

where m is the number of functional nucleotides one would expect in a selection of size n, if selection was random. The standard deviation of m is given by:

$$\sigma = \sqrt{\frac{nM(N-M)(N-n)}{N^2(N-n)}} \tag{7}$$

If the actual observed number of functional nucleotides in selected set is k, we

can compute the z-score of overlap z_o as:

$$z_o = \frac{k - m}{\sigma} \tag{8}$$

Finally, in practice, we calculate both clustering and overlap z-scores over the entire range of ET ranks. We cumulatively bin nucleotides according to their ET rank, then measure the z-scores in each bin. As we are interested in top-ranked nucleotides, we average the z-scores in bins between 0 and 35% rank percentile (0 to 35% ET coverage), to get a single measure, $z_c^{35\%}$ or $z_o^{35\%}$. Note also that the maximum possible number of unique ranks and rank bins is L, the length of the query sequence. However, multiple nucleotides can share the same rank, which leaves a number of unique rank bins empty (not assigned to any nucleotides). We still incorporate these bins into the cumulative measure, by implicitly assigning to them the z-score from the closest valid bin.

Receiver Operating Characteristic Curve

In addition to z_o, we will also provide the area under the Receiver Operating Characteristic (ROC) curve as a measure of ET accuracy. ROC curve is a standard tool for reporting accuracy of binary classification algorithms, and we use it here to provide redundancy.

The goal of the ROC curve is to test whether ET assigns high scores to nucleotides we know *apriori* to be functional (our 'gold standard' reference set). To construct the ROC curve for the ET analysis of a given molecule, we first sort the nucleotides in the molecule according to their ET score, from most to least

important. We then traverse the sorted list in the direction of decreasing ET, and treat each ET unique score as a threshold that determines if a nucleotide is functional. Nucleotides ranked within the threshold are positive predictions P (predicted as functional), while nucleotides below the threshold are considered negative predictions N (predicted as devoid of function). We measure the accuracy of this classification at each threshold level, by calculating the True Positive Rate (TRP) and the False Positive Rate (FPR):

$$\begin{aligned} TPR &= TP/P = TP/(TP + FP) \\ FPR &= FP/N = FP/(FP + TN) \end{aligned} \tag{9}$$

where true positives (TP) are the nucleotides within the threshold that are correctly predicted as functional (they are in the 'gold standard' reference set), while False Positives (FP) are all nucleotides above the threshold that turned out to actually be negative. As we move through the list and evaluate each ET threshold, we plot TRP on the y-axis, and FPR on the x-axis.

If ET is robust at discriminating between functional and non-functional nucleotides, the TPR will increase at a faster rate than FPR, resulting in a ROC curve that is biased toward the upper left corner of the plot (where TPR = 1, and FPR = 0). A perfect predictor, then, will have an area under the ROC curve of 1. A predictor that lacks discriminatory power (positives and negatives receive similar scores) will produce a curve with AUC of 0.5. We use the area under the ROC curve (AUC) as a redundant evaluation metric in this chapter.

2.2.2 Measuring ET Smoothness

In addition to quantifying ET clustering as clustering z-score, we also defined a global measure of clustering we refer to as ET smoothness, SMT. SMT reflects how smoothly ET ranks are distributed over the structure by tallying the rank difference of neighboring nucleotides:

$$SMT = \sum_{i,j} A(i,j)(x_i - x_j)^2 \tag{10}$$

where A is the adjacency matrix as described prior, and x is the ET rank of the nucleotides. In the original work addressing smoothnes [56], we established that evolution tends to minimize difference in evolutionary importance between neighboring residues, because residues exert selective pressure on each other.

2.2.3 Rfam Test Set of Homologous RNA Families

We traced seed alignments of 1070 families from the Rfam database, each family with a minimum of 10 unique canonical sequences (Fig 2.2A). Of these, 71 families with available high-resolution structures made up the structured test set that we tested for ET three-dimensional clustering (Fig 2.2B). Additionally, for a set of 15 families, we compiled a golden standard to test for overlap with ET nucleotides.

2.2.4 ET Code Availability

Evolutionary Trace code, compiled as a command-line utility, along with an example is available at `https://github.com/LichtargeLab/RNA_ET_ms`.

2.3 Results and Discussion

2.3.1 Case study 1: Hammerhead Ribozyme

A first test of ET was the hammerhead ribozyme, a cis-cleaving structure most commonly found in plant viruses, which participates in rolling circle replication by cleaving the nascent transcript [12,58]. The hammerhead motif is not confined to viruses, and new members of the family were recently described in bacteria and eukaryotes [12], where they may support tRNA and siRNA processing, ORF remodeling, and RNAi inhibition. The full-length hammerhead sequence is 60 nucleotides long, and the structure is defined by three short helices that meet at a junction (Fig 2.3A, PDBID 2QUS [59]). There are two main functional domains, the catalytic core that straddles the three-way junction and is responsible for cleavage, and a distal region defined by stem I-stem II tetraloops interactions, which promote efficient folding [60]. Nucleotides composing these two domains are labeled in Fig 2.3A. An evolutionary trace was computed on 26 non-redundant aligned sequences of class I and class II hammerhead ribozymes that represented the major branches of the plant viruses. This led to normalized ET rankings of each base position, from 0% (most evolutionary important) to 100% (least important), that were mapped on the structure in Fig 2.3A. This mapping highlights two clusters of ET bases defined by ET rank percentile below 35% (hereafter referred to as ET bases or ET nucleotides). The first of these clusters consists of 12 ET nucleotides and overlaps the catalytic junction. The second cluster consists of 5 ET nucleotides and overlaps the distal tetraloop region. The functional

nucleotides and their respective domains are listed in Fig 2.3A.

Because clustering by ET residues is a defining feature of aminoacid ET, we assess the clustering of ET nucleotides in the hammerhead. Briefly, to quantify clustering of ET nucleotides, we calculate the number of pairwise structural contacts (distance of 4Å or less) between ET nucleotides, ω, and compare it to the number of contacts formed by same number of nucleotides selected randomly, $\langle\omega\rangle$. Using the standard deviation associated with the random selection, σ, we then express the significance of clustering by ET nucleotides as a z-score $z_c = (\omega - \langle\omega\rangle)/\sigma$. Quantitatively, the clustering z-score z_c is the number of standard deviations that separates the observed number of ET base contacts from the number of contacts expected randomly, and z-scores 2 and above denote statistical significance. See Methods for details.

Using this metric, we calculated clustering of hammerhead nucleotides, binned cumulatively according to their ET rank. We calculated z_c for every bin between 0 to 35% ET coverage (rank percentile), and found that ET nucleotides cluster with a mean z-score $z_c^{35\%} = 3.9$ (Fig 2.3B). Not surprisingly, the clustering profile in Fig 2.3B is similar to the behavior of ET residues in proteins seen in our previous work [55]. We observed a high initial z-score, indicative of ET bases clustering to, hypothetically, form a major functional site, followed by a smooth decline as we expand our ET coverage to include lower-ranked nucleotides. These data confirm that ET nucleotides cluster in the structure, a behavior we would not expect if nucleotide selection was random.

Next, we assessed the second major property of ET: the overlap between ET

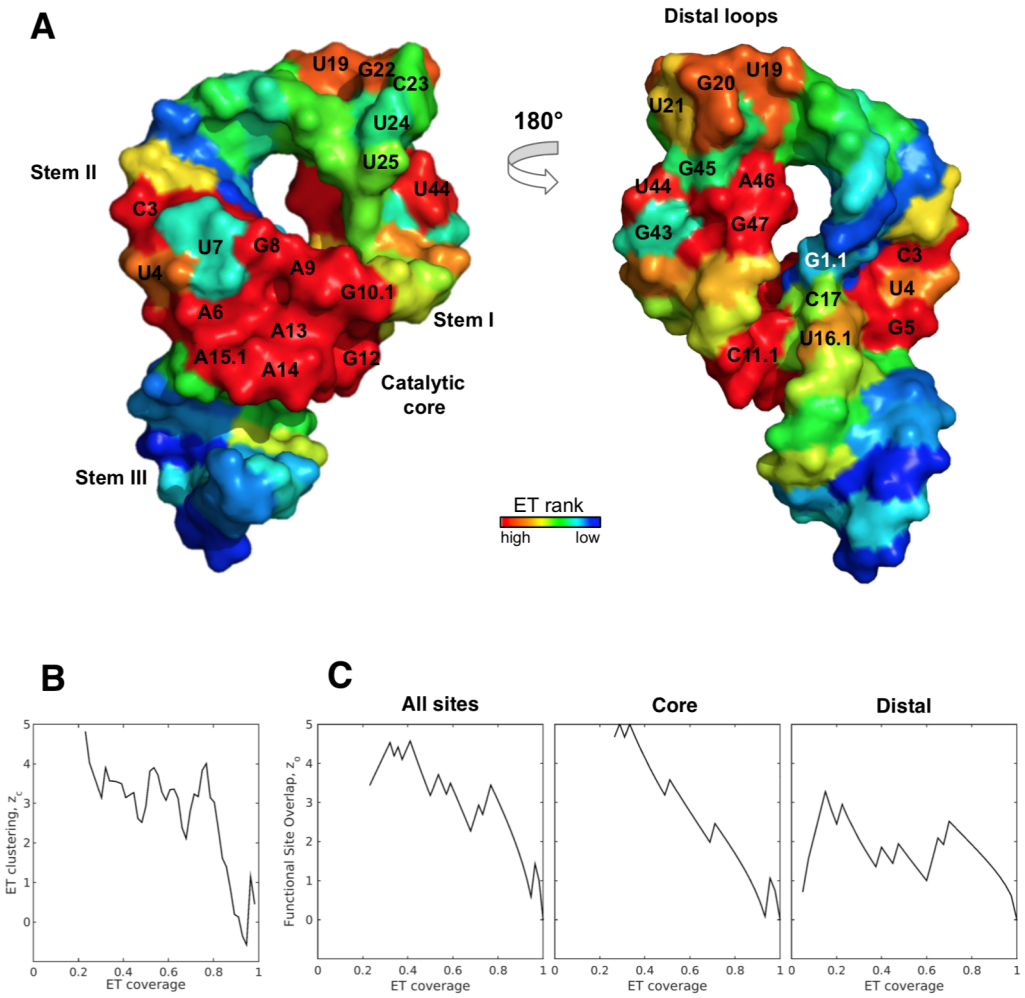

Figure 2.3: ET nucleotides cluster, and predict functional sites in the hammerhead ribozyme. (A) ET ranks mapped onto the structure of the hammerhead reveal clusters of ET nucleotide that overlap canonical functional sites (labeled nucleotides). (B) Clustering by ET nucleotides is statistically significant. (C) ET overlap z-score confirms that ET nucleotides inform both of the canonical functional sites. (To calculate site-specific $z_o^{35\%}$, we remove other known sites from consideration.)

nucleotides and the molecules functional sites. To measure statistical significance of overlap in the hammerhead, we counted the number of active site bases in each ET bin, k, and compared this to the number of active site bases one expects to recover if selecting randomly, m. The random selection was modeled as a hypergeometric distribution, and had an associated standard deviation, which allowed us to convert k into a z-score of overlap z_o (see Methods for details). We calculated overlap z-score for the hammerhead as a function of ET coverage, and it is shown in Fig 2.3C. Once again, note overlap z-score peaks in the 0-35% range indicating that ET nucleotides overlapped strongly with the catalytic core of the hammerhead (mean z-score $z_c^{35\%} = 3.6$). In addition to overlap z-score, we also measured quality of ET prediction in a more conventional manner using receiver-operator-characteristic (ROC) curves, and ET ranking recovers hammerhead active sites with AUC=0.82 (Fig 2.4A). These data show that that ET nucleotides overlap hammerheads two active sites.

Next, we examined the ET clusters in greater detail. The core 12-nucleotide ET cluster overlaps the 16-nucleotide catalytic site of the hammerhead [34] with a site-specific mean z-score $z_o^{35\%} = 4.5$ and AUC of 0.87 (Fig 2.3C center panel, and Fig 2.4B). Notably, the ET cluster contains the key catalytic nucleotides G12 and G8 (both with ET rank percentile of 23%), which are the general base and acid in the cleavage reaction. The ET cluster is also enriched with thermodynamically costly unpaired nucleotides (A6, G5, U4), and nucleotides paired in the non-canonical Hoogsten fashion (C3 forms a Hoogsten pair with U7 and G8, C17 with A13 and U16.1). The highly ranked bases presumably fulfill a critical functional

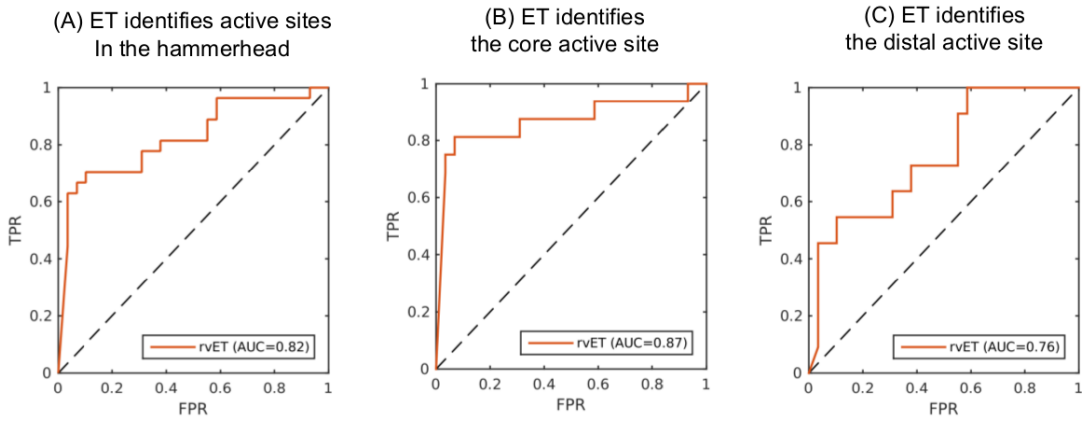

Figure 2.4: ET identifies functional sites in the hammerhead (ROC). ROC AUC measure of prediction accuracy is in agreement with ET overlap z-score. In (A) all active sites are evaluated, while in (B) and (C) the conserved catalytic core and the distal loops are evaluated individually.

role maintained during evolution, leading to their top ET rankings. Critically, mutations in 11 of the 12 nucleotides in this ET cluster completely abolished cleavage activity [34]. These data confirm that the ET cluster directly overlaps the main functional site of the hammerhead.

Of note, four bases in the catalytic core (U16.1, C17, U7, and G1.1) are not part of the 12-nucleotide ET cluster, because their ET rankings of 36%, 52%, 77%, and 96% fall below the 35% threshold we used to define the ET cluster. While U16.1 straddles the 35% threshold and in practice would be considered part of the ET cluster, the three other bases bear closer examination. Their significantly lower ranks suggest that these positions may be under lesser functional pressure, or possibly that there is inherent functional resilience to mutations. To test this hypothesis, we examined carefully the catalytic mechanism of the hammerhead, and the functional role of these three nucleotides.

Although the two nucleotides G1.1 and C17 (ET rank percentile = 96% and 52%) directly participate in the reaction (C17 is the nucleophile and its activated 2-hydroxy group attacks the phosphate group of sessile G1.1 [59]), the critical electron transfer path is along the sugar-phosphate backbone and not the nitrogenous base. As a result, neither the G1.1 nor the C17 nucleotide is under heavy selective pressure. Indeed, direct mutational studies showed that position 1.1 accommodate all four bases, and C17 could also accommodate guanine and adenine (20% reduction in activity [34]). Nevertheless, C17 forms a non-canonical Hoogsten pair with core nucleotide A13, and, perhaps as a result, cannot accommodate uracil (500-fold reduction in activity [34]). In keeping with this greater selective

pressure, C17 has a substantially better ET rank (of 52%) than the sessile G1.1 (96%).

The other notable exception in the catalytic core is U7 (ET rank percentile = 77%). This base is nested among the 12 ET nucleotides, incongruent with its apparent mutational freedom. Yet, the exhaustive mutagenesis studies confirmed that U7 tolerates substitution [34]. While substituting any of the 12 ET nucleotides in the catalytic core reduces activity from 10- to 1000-fold, U7 mutations have no impact on the reaction rate. Thus U7 base identity is not structurally or functionally critical, in keeping with lower ET rank.

In contrast to these data, there is evidence that the last exception, U16.1 which straddles the ET threshold (ET rank = 36%), is critically functional. This base is positioned closely to core nucleotides G12 and C17 (general base and nucleophile in the cleavage reaction), and a recent study suggested that U16.1 could be responsible for coordinating Mg^{2+} in a binding pocket formed by the three bases [61]. The predicted role of the ion is to lower the pKa of G12 to make it more reactive toward C17. Therefore, unlike the three low-ranked exceptions, U16.1 is probably functional, as reflected by its near-threshold rank. Together these data show that the ET ranks of the core catalytic nucleotides are remarkably consistent with the mutational and biochemical interpretation of their functional role.

Next, we examined the apical cluster, formed by the five ET nucleotides (U19, G20, G22, U44, and A46) in the stem I and II loops. This ET cluster overlaps hammerheads11-base tetraloop-tetraloop domain (labeled as Distal region in Fig

2.3A), which is important for the efficient folding of the hammerhead core [60]. Within the domain, ET nucleotides form the structurally critical links between the two stems: U19 of stem I forms a pair with A46 of Stem II, while G20 and G22 bond with G45. Notably, these interactions are the energetically unfavorable Watson-Crick/Hoogsteen pairs, suggesting that hammerhead maintains them through evolution because they are functional. The last ET nucleotide in this cluster, U44, does not form cross-stem interactions, suggesting it serves a different structural role. Of the remaining six (non-ET) bases in the domain, only two form non-canonical interactions, and one of them (U21) is ranked just under the ET threshold (ET rank percentile = 39%). These data show that ET discriminates between the more and less important nucleotides in this domain. Overall, ET bases overlap with the distal domain with a mean z-score $z_o^{35\%} = 2.0$, and ET predicts this domain with AUC of 0.76 (Fig 2.3C right, and Fig 2.4C).

Notably, the two ET clusters are in accordance with the accepted two-step model for hammerhead folding. In the model, stem I and II of the distal region fold first, thereby promoting efficient folding of the catalytic core [62]. The catalytic core is universally conserved, and indeed ET performs very similar to sequence conservation when identifying it, as shown in Fig 2.5A (z-score of overlap) and Fig 2.5B (ROC curve).

The distal regions, however, lacks obvious sequence conservation. As a result its discovery was delayed by several years because researchers focused exclusively on the conserved catalytic core. Ultimately, kinetic and chimeric studies in the full-length hammerhead [60, 62] revealed that tetraloops are a functionally important

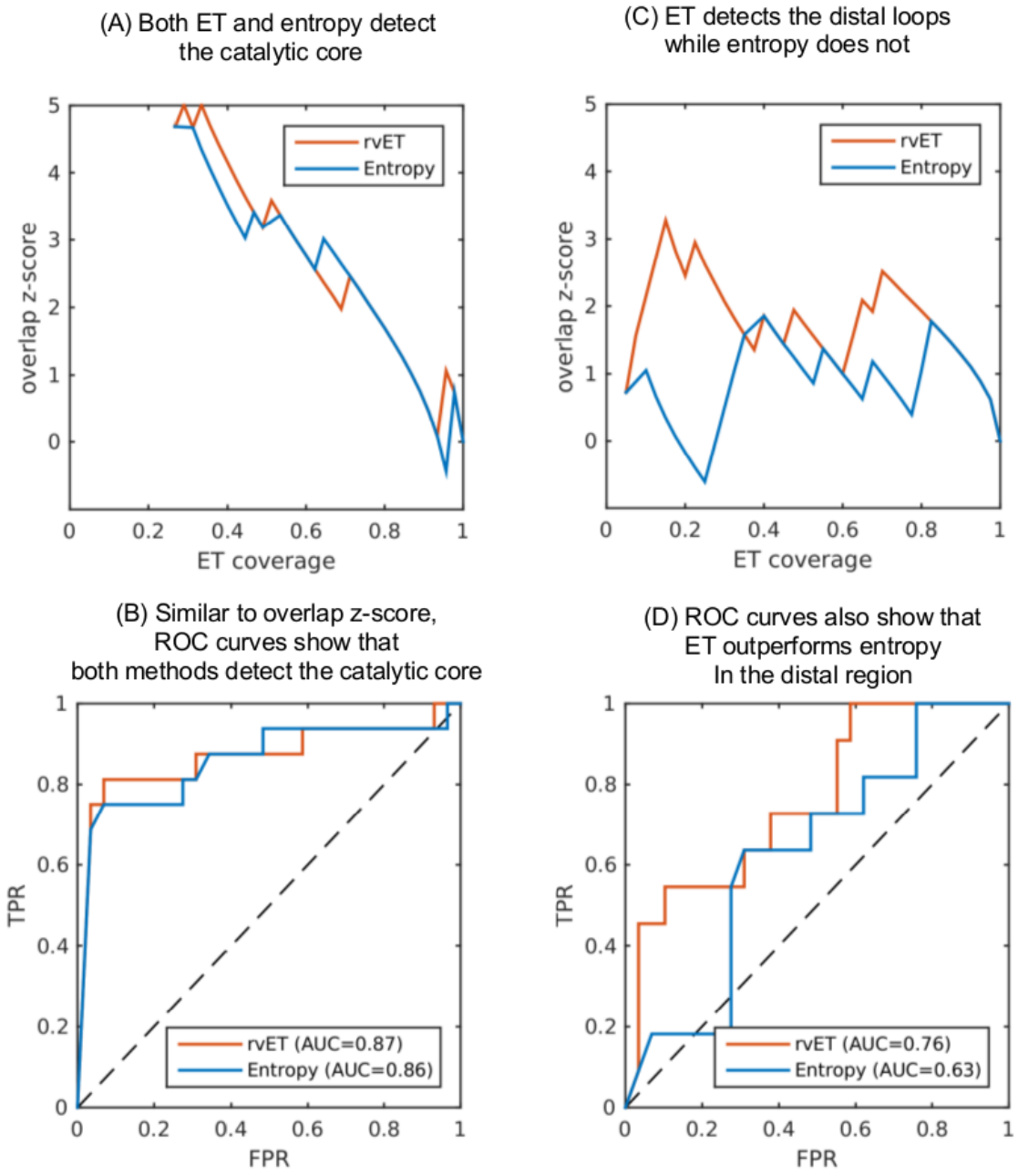

Figure 2.5: ET outperforms conservation in detecting the distal loops. Overlap z-score in (A) shows that both ET and entropy identify the conserved catalytic core of the molecule. However, only ET identifies the distal region, which lacks obvious conservation (B). Represented as ROC curves in (C) and (D), the data support the same conclusion.

domain. Ranking bases according to sequence conservation fails to detect the tetraloop domain, and ET outperforms conservation both in the z-score measure (Fig 2.5C, mean overlap z-score $z_o^{35\%} = 2.0$ vs 0.41), and the ROC AUC (Fig 2.5D, AUC=0.76 vs AUC=0.63). ET detects the distal tetraloops, because while they are fairly variable across the entire tree, they are conserved within their respective class I and class II branches. ET detects this pattern of base variation, resulting in greater predictive power. For comparison, conservation scores are mapped onto the structure in Fig 2.6.

In summary, these data show that ET detects clusters of evolutionary-important bases that define functional domains of the hammerhead. The equatorial ET cluster overlaps the catalytic core of the hammerhead, and the apical ET cluster overlaps the most important bases in the tetraloop domain. Notably, ET outperforms conservation when scoring bases in the tetraloop domain, because it takes into account the evolutionary history of the molecule. Furthermore, where biochemical evidence is available for individual nucleotides, it is remarkably coherent with the nucleotide ranks assigned by ET.

2.3.2 Case study 2: Bacterial Ribosome

The second RNA model system to test ET was the bacterial ribosome. Ribosome is the universally-conserved ribonucleic complex that synthesizes polypeptides from (m)RNA templates. Mature bacterial ribosomes are comprised of two RNA molecules, the 16S and 23S ribosomal (r)RNA, as well as over 50 ribosomal proteins that bind the rRNA during assembly. The RNA component of the ribo-

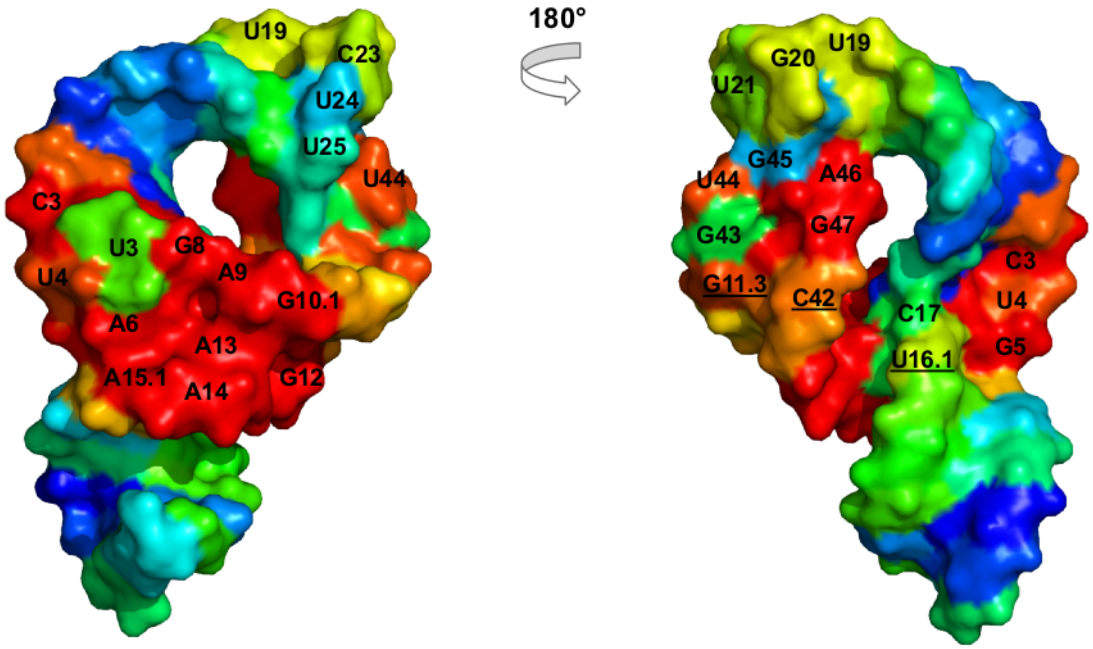

Figure 2.6: Nucleotide conservation in the hammerhead. Conservation scores are produced by Shannon Information Entropy. Note that compared to ET in Fig 2.3A, the distal region of the hammerhead is not nearly as well defined by conservation.

some was originally thought to play a primarily structural role, but high-resolution crystal structures revealed that rRNAs are, in fact, responsible for the catalytic activity of the ribosome. The 16S rRNA decodes the mRNA message by selectively binding acylated tRNAs [63], while the 23S rRNA catalyzes peptide bond formation [6]. Not surprisingly, the ribosome is an important drug target, with over 50 bacterial antibiotics developed to date [25].

Traces for the 16S and 23S RNA were computed using curated alignments of bacterial rRNA provided as part of [64]. ET ranks, mapped onto the three-dimensional structure of the RNAs (Fig 2.7A), revealed that ET nucleotides clustered in the structure (mean ET clustering z-score $z_c^{35\%} = 24.3$ in the 16S, and 32.8 in the 23S, Fig 2.8). Furthermore, ET nucleotides broadly overlapped major functional sites in the ribosome the peptidyl-transferase center in 23S, and the decoding center in 16S (relevant nucleotides marked in Fig 2.7B and 2.7C). Quantification of overlap for these and other major functional sites is summarized in Fig 2.9 for the 16S rRNA, and Fig 2.10 for the 23S (see Fig 2.11 and 2.12 for corresponding ROC curves). In detail, in the 16S rRNA, ET bases overlapped the decoding center ($z_o^{35\%} = 3.9$, AUC=0.91), the tRNA E-, A-, and P-sites ($z_o^{35\%} = 3.5$, AUC=0.88), as well as the mRNA channel (mean ET overlap z-score $z_o^{35\%} = 6.7$, AUC=0.98). In the 23S rRNA, the ET cluster overlapped the peptidyl-transferase center ($z_o^{35\%} = 8.7$, AUC=0.94), the tRNA binding sites ($z_o^{35\%} = 7.6$, and AUC = 0.86), as well as the GTPase-associated center ($z_o^{35\%} = 3.2$, AUC=0.72) and the sarcin-ricin loop ($z_o^{35\%} = 5.3$, AUC=0.85). These data show that ET detects the critically-important active sites in the ribosome.

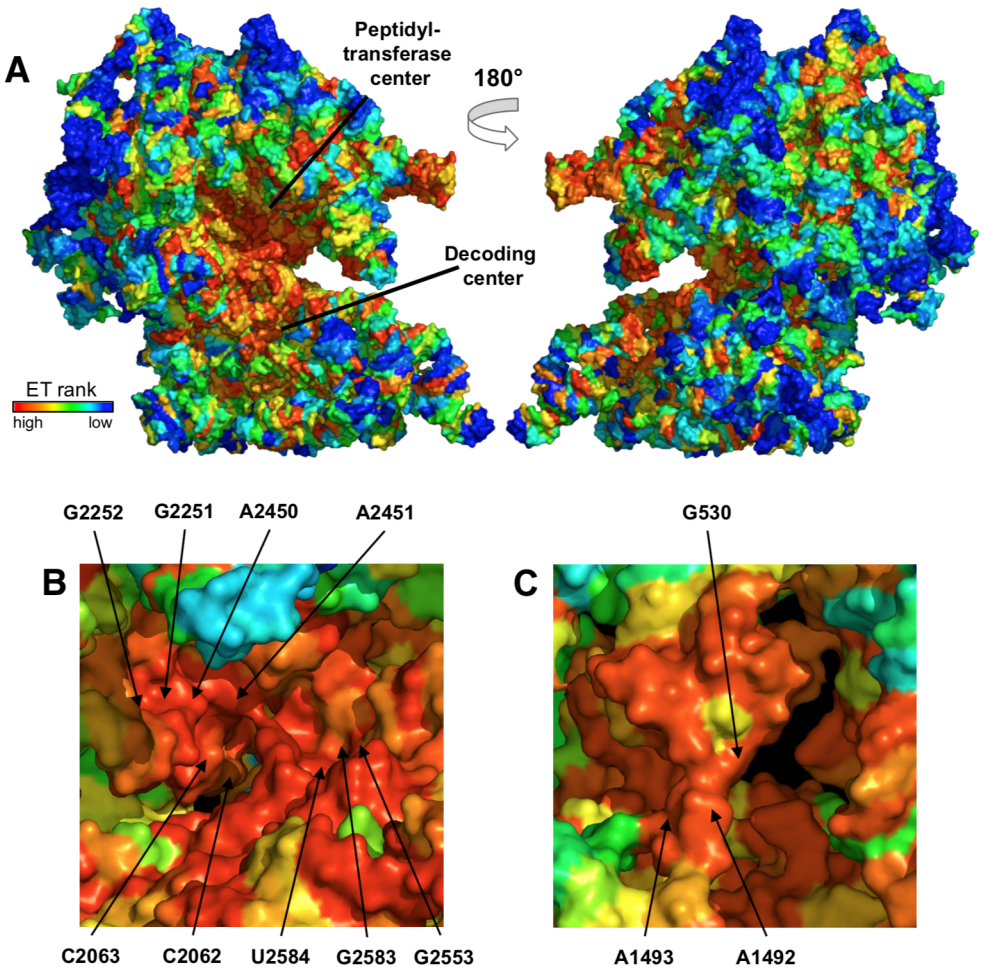

Figure 2.7: ET mapping reveals clusters that overlap active sites in the ribosome. ET ranks are mapped onto the structure in (A). Note the continuous ET cluster that spans both subunits and houses the peptidyl-transferase center (PTC) and the decoding center. Both are shown in detail in (B) and (C), where known catalytic nucleotides are labeled. Structures used are from PDBs 2WDK and 2WDL [65].

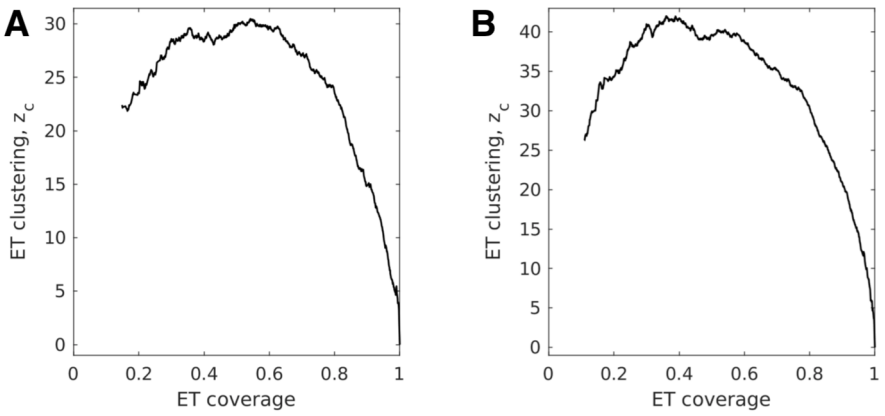

Figure 2.8: ET nucleotides cluster in the ribosome. Clustering z-score z_c for 16S in (A) and 23S rRNA in (B). The high clustering z-scores are indicative of large functional cores.

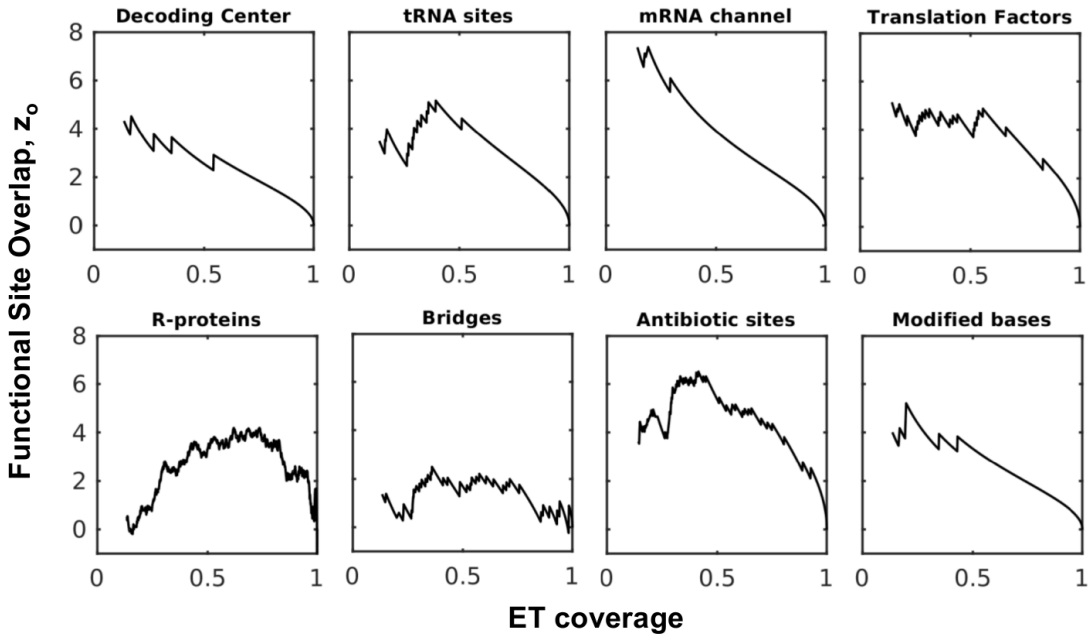

Figure 2.9: ET predicts functional sites in the 16S rRNA. Overlap between ET bases and functional sites in 16S RNA is significant. The sites are decoding center (DC), tRNA binding sites, mRNA channel, translation factor binding sites, structural protein contact sites, antibiotic binding sites, and modified bases.Corresponding ROC AUC are shown in Fig 2.11.

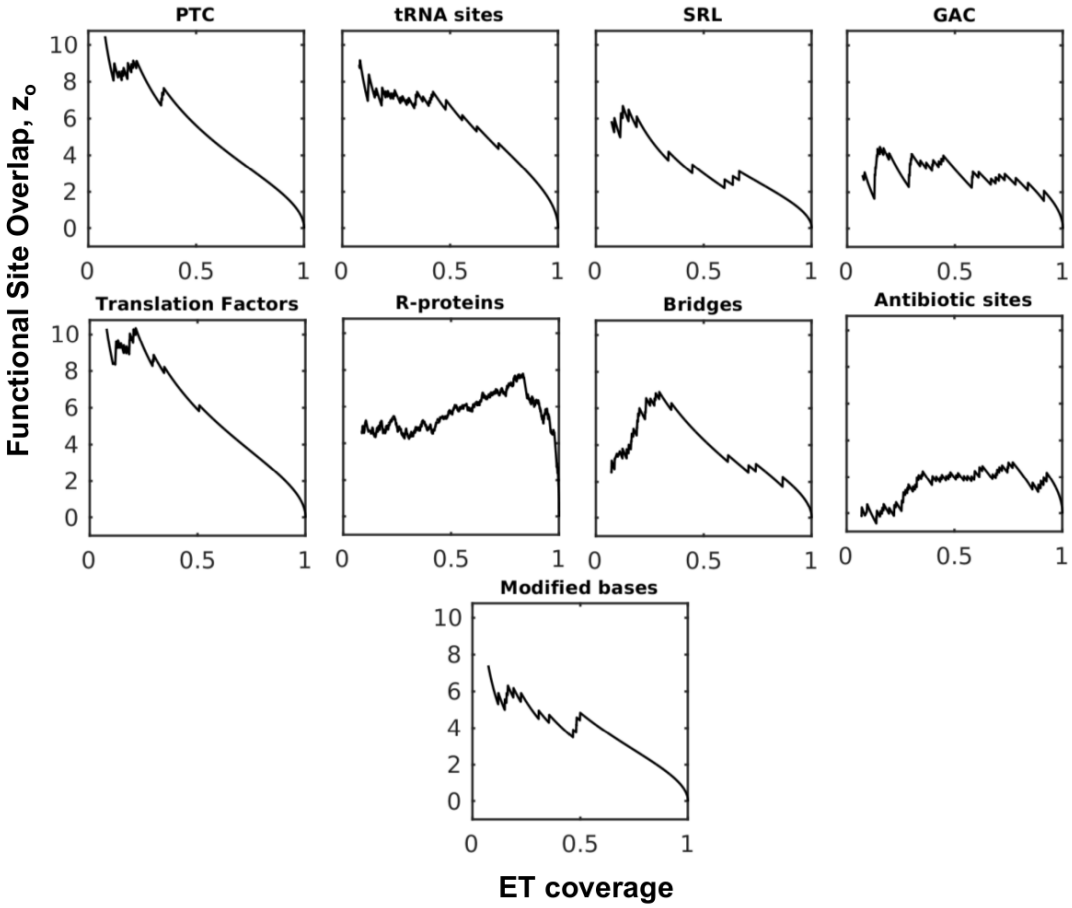

Figure 2.10: ET predicts functional sites in the 23S rRNA. Overlap between ET bases and functional sites in 23S RNA is significant. The sites are peptidyl-transferase center (PTC), tRNA binding sites, sarcin-ricin loop (SRL), GTPase-associated center (GAC), translation factor binding sites, structural protein contact sites, antibiotic binding sites, and modified bases. Corresponding ROC AUC are shown in Fig 2.12.

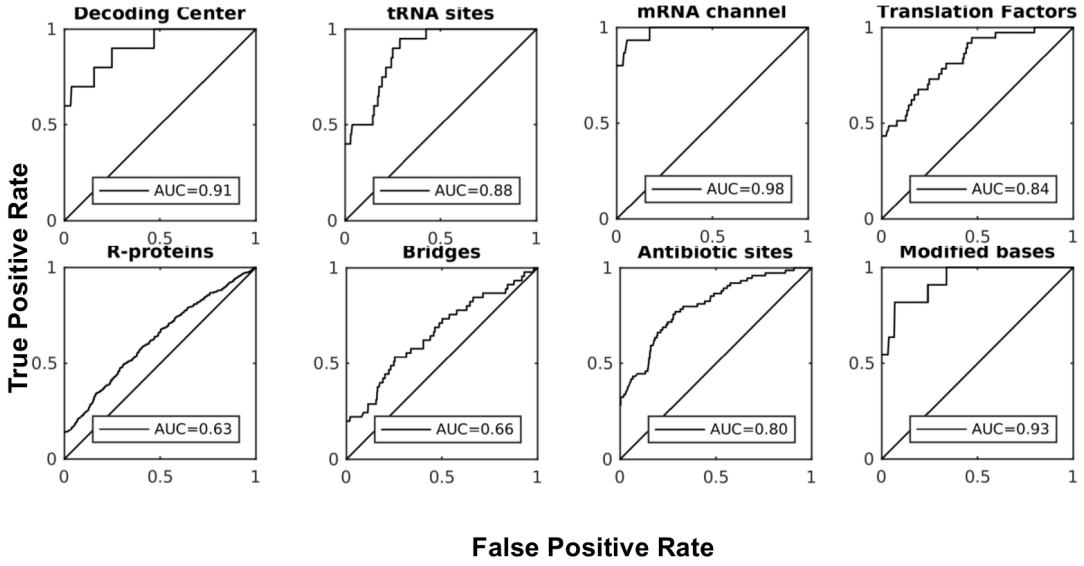

Figure 2.11: ET predicts functional sites in the 16S rRNA (AUC). These data closely agree with z-scores in Fig 2.9.

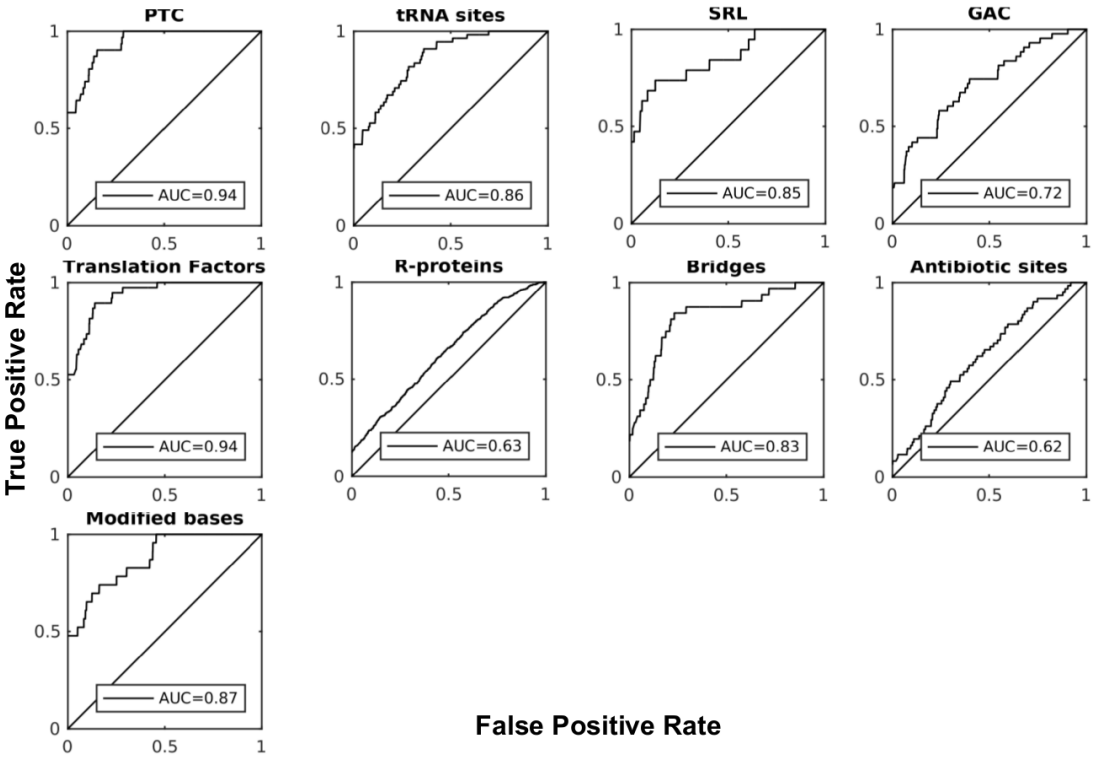

Figure 2.12: ET predicts functional sites in the 23S rRNA (AUC). These data closely agree with z-scores in Fig 2.10.

Next, we tested whether ET bases also overlap protein binding sites (refer to panels in 2.9 and 2.10). First, we examined the binding sites of bacterial translation factors (including IF1, EF-Tu, EF-G, RF1, and RF3 for which high-resolution structures are available). They transiently bind the ribosome and are required for timely initiation, elongation and termination of translation. We defined binding sites as all rRNA bases within 4 of the protein, and found that ET bases overlap TF binding sites with mean ET z-score $z_o^{35\%} = 4.8$ in the 16S rRNA and 9.3 in the 23S (AUC=0.84 and 0.94 respectively). Next, we tested the binding sites of structural ribosomal proteins, which serve as scaffolding and are enfolded by the rRNA during assembly [66]. ET bases overlap these contact sites with mean ET z-score $z_o^{35\%} = 0.96$ and 4.7 in 16S and 23S respectively (AUC = 0.62 and 0.63). As expected, overlap between ET bases and translation factor binding sites is markedly higher than the ET overlap with structural protein contact sites. This is in line with the expectation that structural proteins are not as critical to function. Unexpected is the discrepancy between ET overlap for r-protein sites in 16S and 23S rRNA ($z_o^{35\%}$ of 0.96 vs 4.7). Interestingly, r-proteins in the 16S occupy 42% of all nucleotides, compared to only 29% in the 23S rRNA. This implies higher specificity of binding in the 23S subunit, resulting in better overlap with ET nucleotides. These data reflect ETs sensitivity to the evolutionary pressure exerted across the ribosome. ET clearly separates more important sites, such as the catalytic core and TF binding sites, from the r-protein sites. In summary, ET ranks of the nucleotide correlate strongly with their functional impact.

Next, we also tested the critical structural bridges [67] that connect the two

subunits for overlap with ET nucleotides. We found that ET bases overlap the bridges in the 23S ($z_o^{35\%} = 4.5$ and AUC=0.82), but not the 16S subunit ($z_o^{35\%} = 1.2$, AUC=0.66). To explain this difference, we examined the molecular basis of the bridge contacts, and found that 22 out of 48 contacts on the 16S side are formed by the nucleotide phosphate backbone, compared to only 7 out of 30 on the 23S (the difference is significant with hypergeometric p-value of 0.03). Because phosphate contacts are non-specific (not depended on the identity of the base), the nucleotides in 16S bridges are not evolutionary constrained and are ranked lower by ET. These data show that while ET is able to detect critical structural elements, the underlying molecular determinants can produce exceptions to the ET model, similar to the earlier example of the catalytic mechanism in the hammerhead.

Because ET bases broadly overlap functional sites, we next asked: do known ribosomal antibiotics also target ET bases? To determine if ET recovers known antibiotic binding sites, we compiled a list of binding sites for 32 different antibiotics [11]. We quantified recovery of these sites by ET, and found that ET bases overlap antibiotic binding sites with mean ET z-score $z_o^{35\%} = 0.41$ in 23S and 4.3 in the 16S (AUC = 0.62 and 0.80). While the data confirm that ET bases overlap strongly with antibiotic binding sites in the 16S rRNA, we asked why there was a lack of overlap in the 23S subunit. We examined the molecular basis for antibiotic action in 16S and 23S rRNA. We found that the 23S antibiotics mainly target the exit channel, which is a tunnel that traverses the subunit, and it is used to extrude the nascent polypeptide from the ribosome. Because the channel does not serve as a site for catalysis or binding, it is lined with nucleotides that are not under heavy

selective pressure. As a result, 23S antibiotic binding sites have a modest overlap with ET bases. In contrast, we found that in the 16S RNA, antibiotic families primarily target the mRNA binding channel and the decoding center, which, as we already showed, are primarily composed of high-ranked ET bases. Thus, the molecular mechanism of antibiotic action is in line with ET nucleotide ranking.

Next, we tested ET's ability to detect bases that do not necessarily belong to an established active site or binding interface, but are nevertheless hypothesized to be important: modified bases and those with known deleterious mutations We first examined modified bases; nascent rRNAs can be post-transcriptionally modified, and at least 34 nucleotides in the ribosome carry modifications [68]. While the exact role of modified bases is unclear, ribosomes assembled from unmodified rRNA are less active than wild type [69]. ET overlaps modified bases with mean ET z-score $z_o^{35\%} = 4.0$ in 16S and 5.9 in the 23S (AUC=0.93 and 0.87), suggesting that these bases evolved to perform a function in the large subunit. In addition to the modified bases, we examined nucleotides with known deleterious mutations. We compiled a list of mutations available on the Comparative RNA Website database [70], and sorted them into unambiguously deleterious or benign. Applying ET, we see that while both cohorts overlap with ET bases, there is clear separation between the two categories of mutations (Fig 2.13). In the 16S, ET bases overlap with deleterious mutations more frequently than with benign mutations (mean z-score $z_o^{35\%} = 4.5$ and AUC=0.92 for deleterious mutations, and $z_o^{35\%} = 1.9$ and AUC=0.69 for benign). In the 23S subunit, the difference is also present (mean z-score $z_o^{35\%} = 6.1$ and AUC=0.94 for deleterious mutations,

and $z_o^{35\%} = 2.9$ and AUC=0.81 for benign). Nucleotides with benign mutations rank consistently lower than nucleotides with lethal mutations. These data further point at a clear connection between evolutionary importance, as measured by ET, and functional impact.

Finally, we examined ET clusters that do not overlap with a known functional site. From the ribosomal structure, we excluded all nucleotides composing known sites, as well as all nucleotides within 10 of a r-protein. This exclusion analysis produced 4 clusters of high-ranked ET nucleotides on the surface of the ribosome, three in the 23S and one in the 16S (Fig 2.14). The ET nucleotides in these clusters (listed for each cluster in Table 2.1) are undocumented in the literature and carry no obvious functional significance. We propose that clusters in the 23S could serve as sites for binding of regulatory proteins and chaperones, or as sites of ribosomal processing during maturation and assembly. Meanwhile, the ET cluster in 16S (Fig 2.14D) is located adjacent to helix h5, which acts as a binding site for several translation factors [71, 72]. It is therefore possible that nucleotides in this cluster are involved in allosteric regulation of translation. These data show that ET-guided structural analysis can suggest sites of interest even in the well-studied systems such as the ribosome.

In summary, we discovered that ET nucleotides cluster on the structure of the ribosome, and that the core ET cluster clearly defines the major functional sites of the molecule. Additionally, ET ranking also suggests protein and antibiotic binding sites. We also show that higher-ranked ET nucleotides are enriched for inactivating mutations and post-transcriptionally modified bases. In detecting

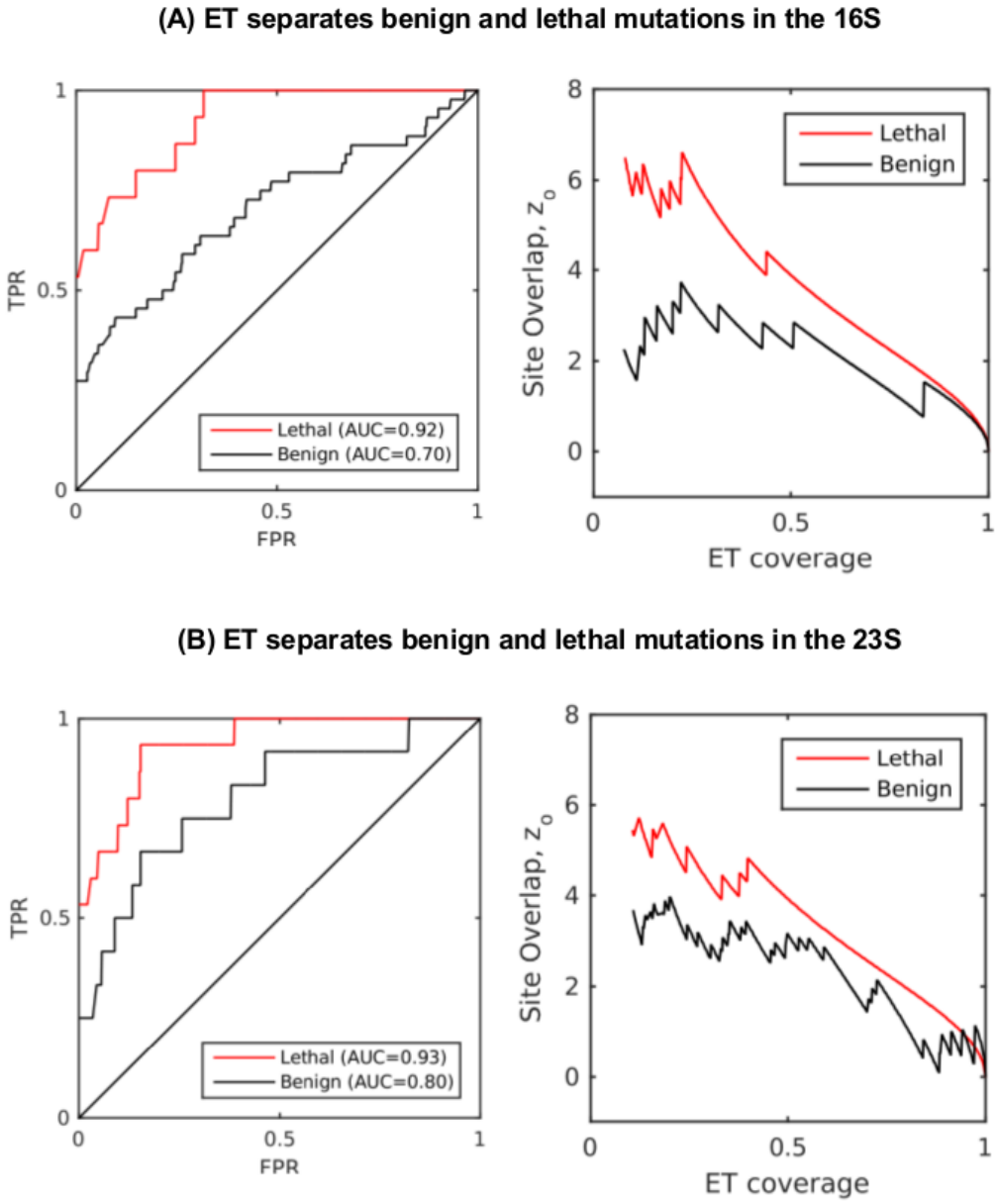

Figure 2.13: ET discriminates between lethal and benign mutations in the ribosome. Sites with benign mutations are ranked as less important than sites than display lethal mutations. This is true in both (A) 16S and (B) 23S rRNA. Overlap z-scores and ROC curves are in agreement.

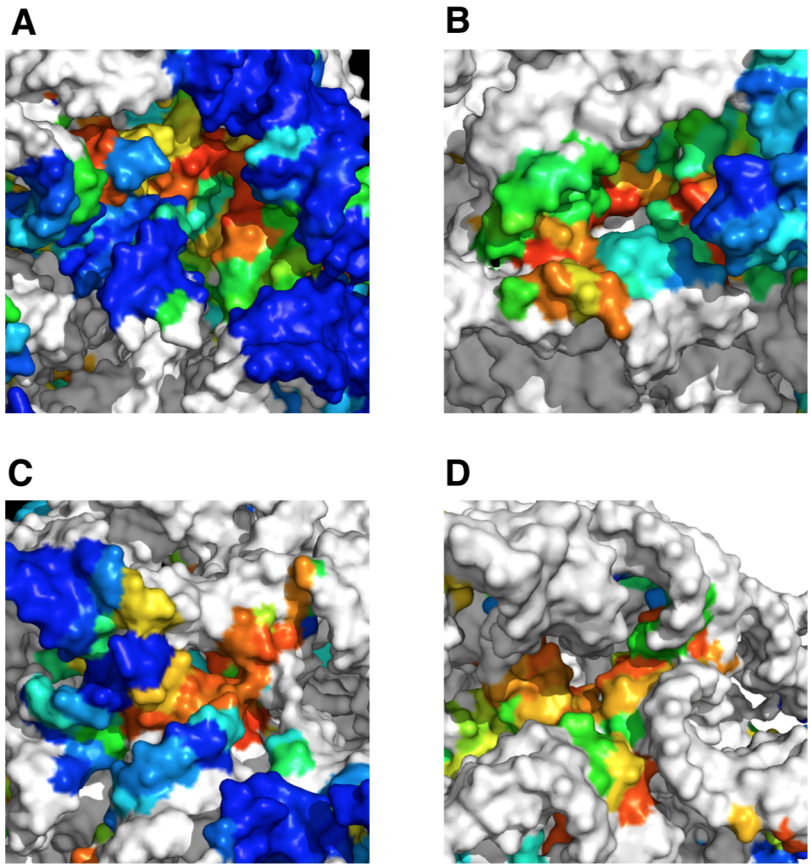

Figure 2.14: Potentially novel ET clusters in the ribosome. Clusters (A-C) in the 23S rRNA could serve as binding sites for regulatory proteins, while cluster (D) in the 16S is close to a translation factor binding site, and therefore could play a role in translation. Nucleotides in white are either known functional sites, or within 10 of a r-protein, and therefore excluded from analysis. Nucleotide composition of these clusters is listed in Table 2.1

.

Cluster	rRNA	Nucleotide in cluster
A	23S	C45, G48, A49, G51, A52, U120, A213, G214, G215, A216, G219, G220, A221, A222, G232, A233, C234, C237, A255, A256, A262, A265, G266, A422, C426, U427, A428, A429, G430, U431, A432
B	23S	G1281, G1283, C1289, C1298, G1299, A1301, U1313, C1314, C1315, U1329, C1330, C1604, C1605, A1608, A1609, A1641, G1642, G1643
C	23S	C32, U33, G35, U306, G307, A454, G473, G474, U475, G476, A502, A503, A505, G506, U511, G1212, A1213, U1234, G1235, G1236, A1237
D	16S	A51, A53, C54, G115, A116, G117, U118, A119, A313, C314, A315, G317, G318, A1468

Table 2.1: Composition of undocumented ET clusters in the ribosome.

these sites, ET is more accurate than conservation (2.15), according to both z-score and ROC AUC. Finally, because these data indicate that the ET model applies to the ribosome, we suggest several potentially novel sites.

2.3.3 Generalizing the Model to Other fRNA families

The hammerhead and ribosome case studies are consistent with two fundamental ET properties: that top-ranked bases cluster structurally, thereby revealing functional sites. To assess the generality of these features, ET was next tested on RNA families in the Rfam database. We selected 1070 RNA families that had at least 10 canonical sequences in their Rfam seed alignment. This set of RNAs included a broad selection of classes, including riboswitches, tRNAs, RNAzymes, viral particles, small regulatory RNA, and lncRNA (Fig 2.2A). Additionally, among these are 71 families that can be paired with at least one high resolution structure. Each alignment was traced, and the trace was evaluated for clustering among ET bases as a function of ET coverage.

We first evaluated the high-resolution structured set, and found that 64 out of 71 RNAs yield clustering z-score $z_c^{35\%}$ greater than 2, indicating that ET is detecting clusters of important nucleotides (Fig 2.16A, white). This set consisted of well-ordered full length structures including riboswitches, RNAse P, catalytic introns, ribozymes,and rRNAs (Fig 2.2B). Notably, ribosomal and splicesomal RNA displayed larger z-scores compared to the rest of the set (14 or more), suggesting large and evolutionary-important core functions. We then examined the 7 Rfam families that did not show structural clustering by ET nucleotides. Three of

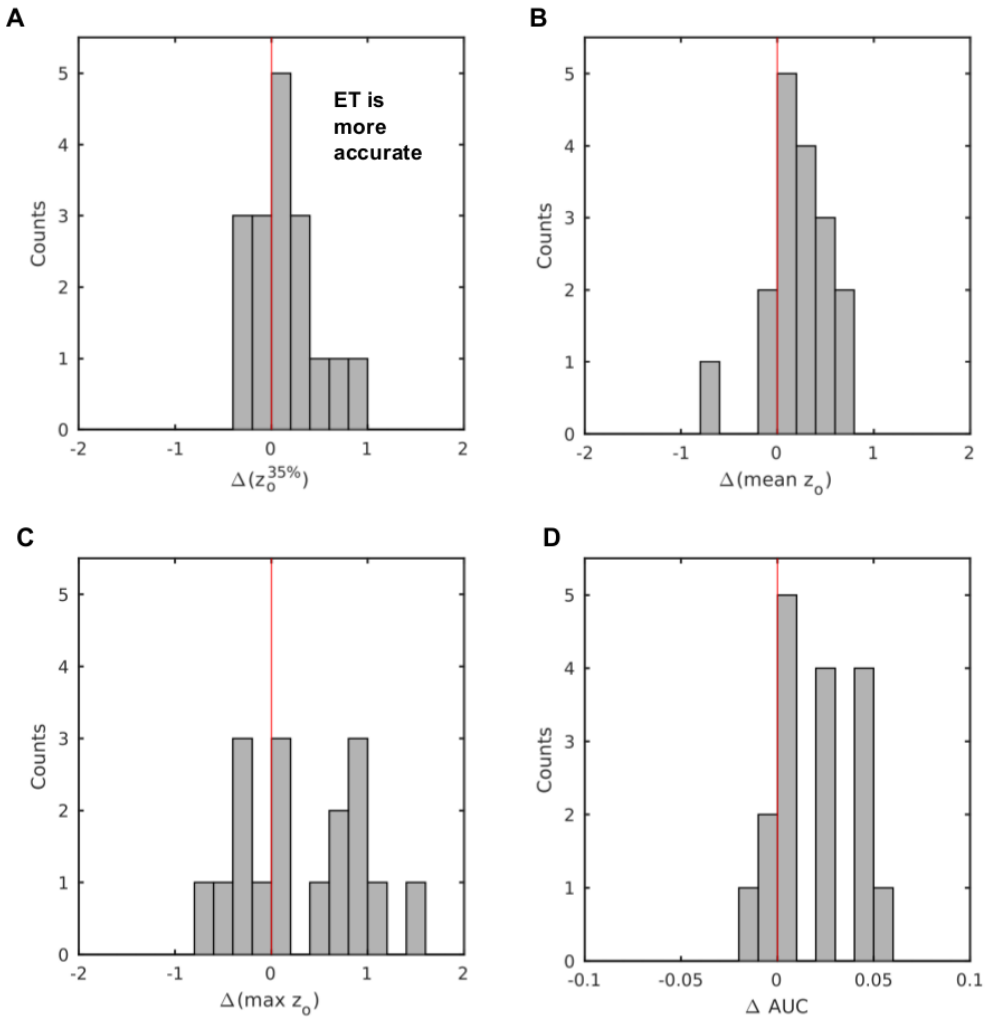

Figure 2.15: ET is more accurate in detecting ribosomal sites than conservation. For each of the 17 ribosomal functional sites in our test set, we measured the difference, Δ, in prediction accuracy between ET and conservation (Shannon Information Entropy). The four metrics of prediction accuracy used are (A) mean z-score of overlap for nucleotides bins ranked in top 0-35%, $z_o^{35\%}$, (B) z-score of overlap averaged over all rank bins, z_o, (C) maximum overlap z-score z_o^{max}, and (D) area under the ROC curve. While the results agree closely, ET outperforms conservation in most cases in all 4 of the prediction metrics.

those families were small viral structures (approximately 30 nucleotides in length) found in the Human Immunodeficiency Virus (HIV) RNA. Their sequence alignments consisted of highly similar sequences (mean sequence identity 91%), and their narrow phylogenetic scope precluded meaningful ET analysis. By contrast, the average mean sequence identity in successful examples was 64%. Finally, in each of the remaining 4 families that performed poorly, clustering could not be fully assessed, because their best matched structures were a fragment of the whole length molecule. Overall, however, these data show that the model is in keeping with observations in 90%of the fncRNAs we were able to test. Failures rarely but consistently associated with missing structural context, or a deficit of evolutionary information due to a lack of sufficiently divergent sequences.

Next, in order to test ET for RNAs without known three-dimensional structures, which is 93% of our test set, ET base clustering was assessed in the primary sequence (1-dimension), and in the secondary structure models provided by Rfam (2-dimensions). 83% of secondary structures (Fig 2.16A, gray) displayed ET clustering with $z_c^{35\%}$ above 2. The few outliers with large z-scores, once again, were rRNA and spliceosome subunits. ET clustering could also be detected in the majority of primary sequences, with 62% reaching $z_c^{35\%}$ of 2 (Fig 2.16A, black). These data show that while secondary and primary structures lack the nuanced three-dimensional context, they nonetheless reveal clusters of ET nucleotides. One possible application of this property is to use ET clustering in predicted secondary structures to distinguish between poor and robust models.

Finally, of the 71 RNA families tested for three-dimensional ET clustering, we

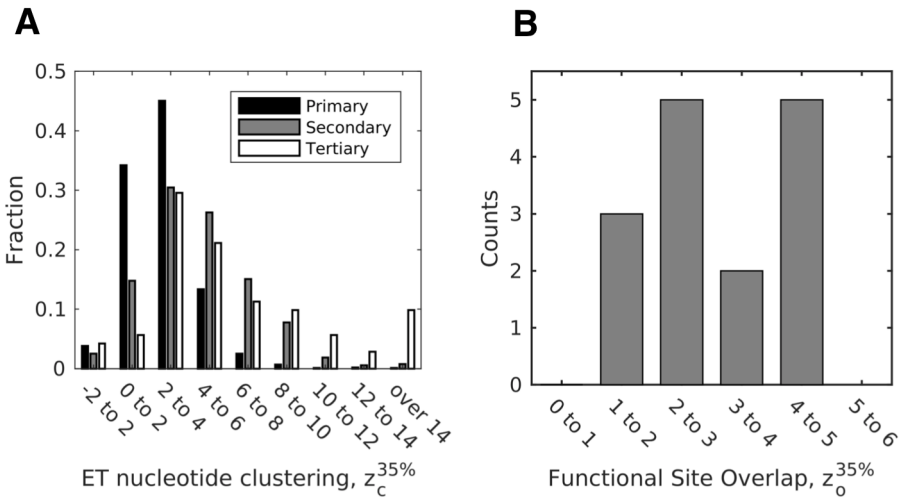

Figure 2.16: Clustering of ET nucleotides, and their overlap with functional sites is general. Z-scores above 2 indicate statistically significant clustering and overlap. Shown in (A) are aggregate ET clustering data for 1070 RNA molecules. Et clustering is detected in 62% of primary sequences, 83% of secondary structures, and 91% of tertiary structures. In a subset of 15 of these molecules (B), we measure overlap with known functional sites. In 12 of 15 test cases, overlap is significant.

selected 15 for functional site analysis. These included the hammerhead ribozyme, the two ribosomal subunits, tRNA, RNAse P, group I self-splicing introns, and 9 riboswitches. For each of these molecules, we searched the literature for the canonical functional sites, and then computed their overlap with ET bases. In 12 of 15 cases, functional site overlap z-score $z_o^{35\%}$ was above 2.0 (Fig 2.16B). In two cases, the THF riboswitch and the PreQ1 riboswitch (RF01831 and RF00522), overlap approached significance with $z_o^{35\%} = 1.86$ and 1.84. Finally, functional site overlap z-score $z_o^{35\%}$ was 1.44 for the FMN riboswitch (RF00522). Interestingly, its seed alignment contained a number of misaligned sequences; removing them, and retracing, raised $z_o^{35\%}$ from 1.44 to 1.9.

Together, these analyses of ET clustering and overlap suggest that the ET model is general and applicable to a wide range of functional RNAs.

2.3.4 Optimizing Sequence Selection Improves Performance

Since, in RNA, ET fulfills the same clustering and functional site overlap properties as in proteins, perhaps that likewise improving the quality of the structural clusters can guide improvements to the quality of functional site predictions? In proteins the two correlate strongly. As a result, improvements in ET clustering can be used to optimize sequence selection, which in turn produces better functional site recovery [55], with important practical ramification in optimization of analyses [73].

To test this hypothesis, we assessed two different metrics of cluster quality: ET clustering z-score, as described earlier, and the ET smoothness. ET smooth-

ness is the cumulative ET rank difference between all neighboring nucleotides in the structure (meaning lower absolute values for smoothness corresponds to a smoother distribution of ranks). This measure reflects smoothness of evolution over the entire structure, and is a more holistic metric than the mean ET base clustering [56].

We tested the relationship between the clustering metrics and the quality of prediction in 15 RNA families with curated functional sites. For each family, we generated a set of 1,000 alignments by randomly shuffling bases in the original alignment. We traced the alignments, and measured their smoothness, and their mean overall clustering and overlap z-scores, z_o and z_c. We then binned the alignments by their shuffle rate, and averaged the scores, as shown in Fig 2.17A) for glmS riboswitch. As seen in the glmS example, as we introduce errors into the alignment, ET overlap, clustering, and smoothness deteriorate in highly correlated manner. Across the 15 test cases, mean correlation between ET overlap and clustering was $r = 0.95$, and $r = -0.96$ between overlap and smoothness (Fig 2.17B).

The correlation between structural quality of the trace and overlap has practical implications, because smoothness and ET clustering can be used as indirect measures of trace fidelity. By optimizing trace alignments to maximize smoothness (or ET clustering), we, presumably, also maximize accuracy of active site prediction. We tested this hypothesis in the nine test riboswitches. For each riboswitch, we generated a starting alignment, distinct from the seed, and based on 500 sequences chosen at random from the familys full sequence repository on Rfam. Us-

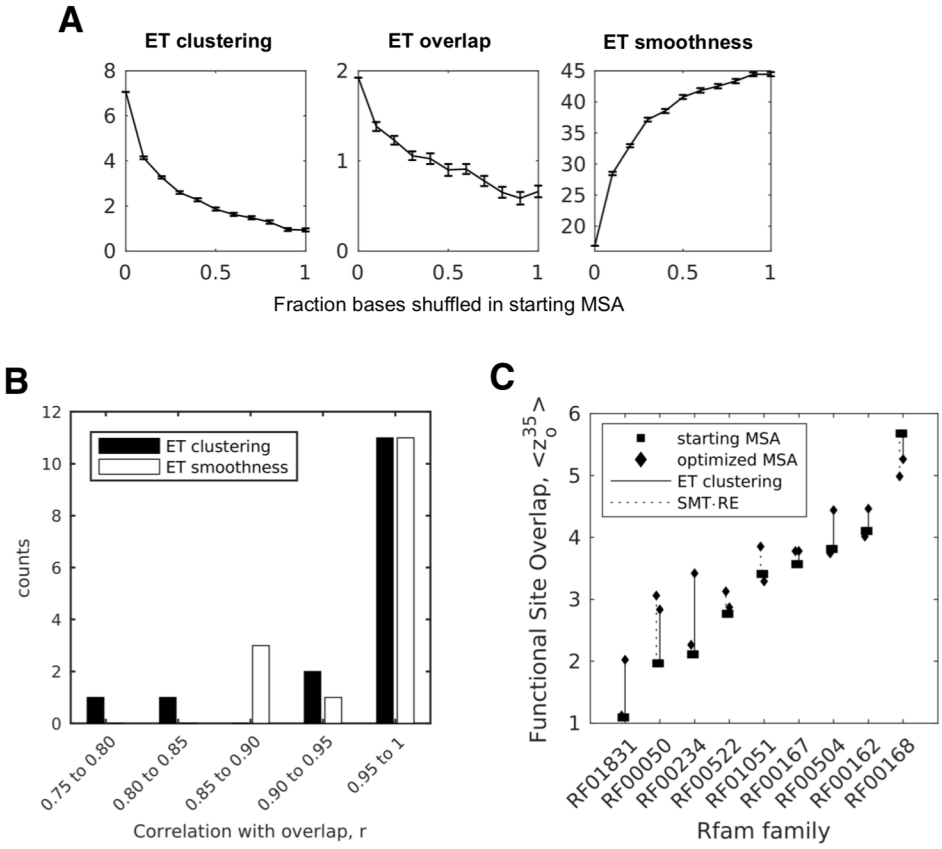

Figure 2.17: Optimization of input alignments via ET clustering and smoothness improve overlap. Degrading the glmS alignment in (A) shows that clustering and smoothness are correlated with overlap. Applying this analysis to 15 molecule shows that the correlations are general (B). This suggests that sequence selection can be optimized to produce a more effective ranking of ET. This is shown in (C), where we use two different clustering measures, to select alignments that produce better functional site predictions than starting alignments.

ing these alignments as reference, we then generated for each riboswitch an ensemble of 5,000 alignments by removing a random number of sequences $0 < n < N20$ (where N is the size of the alignment) from the reference. We then traced each alignment, and computed its ET smoothness, ET clustering z-score $z_c^{35\%}$, and ET overlap z-score $z_o^{35\%}$. Finally, we chose from the random ensemble, the alignment that has the smoothest ET, or the highest ET clustering. We then compared overlap z-scores produced by the smoothness- and clustering-optimized alignments to the original Rfam alignments.

Startlingly, we found that in 7 of 9 cases, optimization via ET smoothness led to decrease in performance, with a mean reduction of 13%. To explain this behavior, we examined the smoothest alignments, and found that the optimization selected alignments consisting of a small number of highly invariant sequences. These alignments sacrificed phylogenetic diversity as a trade-off for a very smooth, yet uninformative, ET ranking of highly conserved nucleotides. To address this problem of narrow phylogenetic scope, we introduced Rank Entropy (RE), $\sum_r^N fr \log fr$, which measures the frequency (fr) of each ET rank (r) in the alignment. RE is maximum when each column is assigned a unique rank, and is zero when the alignment is entirely conserved.

Normalizing RE and ET Smoothness to 1, and then using their product as a single measure (RESMT), rescued optimization performance in 8 of 9 cases, yielding mean improvement of 9% in prediction quality (Fig 2.17C, dashed lines). Finally, we tested mean ET clustering as an optimization metric. We found that maximizing mean ET clustering z-score produced alignments that yielded an aver-

age improvement of 24% (Fig 2.17C, solid line). ET clustering z-score performed better than smoothness-based metrics, because it expresses significance of ET nucleotide clustering relative to the remaining structure. When the conserved ET bin dominates the trace (ET nucleotides are more likely to cluster), this results in a greater expected clustering weight, a wider standard deviation, and as a result a lower clustering z-score, making this measure more sensitive than raw ET smoothness. Together, these data show that optimization of input alignments is both possible and useful.

In summary, sequence selection can be optimized in order to achieve better active site recovery. Input sequences are the principal factor that affects the quality of the trace. Using ET clustering and RE-smoothness as indirect measures, we can remove sequences that are either too phylogenetically distant or erroneous. In this manner, we can elevate trace quality and more accurately predict the active sites.

2.4 Conclusion and Future Direction

Over 50 years ago, Carl Woese and Francis Crick hypothesized that RNA could serve as a precursor to DNA and proteins [3], and since then, RNAs have been found to perform a variety of roles within the cell. Here, we show that the similarity between functional RNAs and proteins is not just a subject of anecdotal likeness. We argue that this similarity is based upon the fundamental principle of selection that governs evolution of function in both RNAs and proteins. In the

same manner as evolutionarily important amino acids in proteins, evolutionarily important RNA nucleotides evolve in compact, non-random clusters that inform on the function of the molecule. As we have shown in a number of examples, including the hammerhead and the ribosome, these clusters correspond to catalytic sites, ligand-binding pockets, and molecular interfaces, and are enriched for inactivating mutations and modified nucleotides. These basic properties underline the relationship between sequence, structure, and function in structured RNAs, and suggest that it is possible to identify novel functional sites in structured RNA molecules using Evolutionary Trace analysis. Furthermore, we demonstrate that there is a quantifiable correlation between cluster formation and recovery of functional sites. This correlation is directly informed by ET clustering and ET smoothness, which measures the evolutionary history similarity in nearby amino acids or nucleotides. In structured RNA, just like in proteins, evolution tends to minimize rank differences between neighboring nucleotides.This leads to formation of smooth clusters that inform function. In practice, this property allows us to maximize prediction accuracy of ET by constructing alignments that maximize spatial clustering and smoothness of ET nucleotides.

As this study has focused on well-known structured RNAs, it is not clear whether ET properties are common to all RNA classes, especially novel classes such as lncRNA which have a tenuous link between sequence, structure, and function. This limitation of ET has partly been demonstrated in the two case studies. ET erroneously flags critical structural nucleotides as unimportant, when the function of the nucleotide is decoupled from its nitrogenous base, as we have

observed in the hammerhead with nucleotides G1.1 and C17, and with the bridge nucleotides in the 16S rRNA. Because the phosphate backbone assumes function in these nucleotides, the bases are unrestrained and do not display the characteristic pattern of branch-specific sequence conservation that ET quantifies as evolutionary-important. Therefore, it is important to be mindful of possible false negatives when applying ET to a poorly understood molecule.

The future directions of this work is three-fold. First, while RNA ET is currently available as a command-line tool, we wish to also create an online-based server to make the method accessible for the general audience. Second, the ET paradigm is general to any evolving polymer, and we would like to extend this work to functional non-protein-coding DNA sequences, for example gene promoters. The current implementation of the tool already supports analysis of DNA sequences, and ET would benefit from explicit validation on DNA. Finally and most importantly, the authors would like to apply ET in relevant experimental work to predict functional sites in fRNA.

In fact, we are currently assisting AR Gener at the Baylor College of Medicine in his work on HIV. Briefly, his central hypothesis is that HIV interferes with human genome remodeling by interacting with host protein CTCF, a known regulator of human genome architecture [74]. To test this hypothesis, he identified 18 putative CTCF binding sites in HIV, and is preparing to assay them for binding to CTCF. To help guide his inquiry, we traced a representative alignment of 3,666 full-length HIV sequences, and ranked each CTCF site according to the average ET score of its nucleotides. We hypothesize that the highest-ranked CTCF sites

are also more likely to bind to CTCF, and these sites will be tested first. While experimental validation is still pending, this application of ET provides an example of its utility in RNA research. We expect the need for such analyses to grow in the future, as the scope of RNA research continuous to expand.

Chapter 3

The Problem of Function Annotation in Biology

In Chapter 2, we explored functional discrimination of individual nucleotides within the confines of a single RNA molecule. In Chapter 4, we address the problem of function determination from a different perspective, that of determining functions of individual proteins within the context of a cell. The basic definition of this problem is rather straight-forward: given all possible proteins (20,000 in human) and all possible functions (approximately 46,000 functional labels codified in the Gene Ontology), we arrive at $20,000 \times 46,000$ 960 million possible protein-function pairings. Within this universe of possible functions, how do we find those that actually exist in the cell? While there are high-throughput experimental techniques developed for such a task (two-hybrid screens [75], protein microarrays [76]), they cannot exhaustively scan all 960 million combinations. Not surprisingly then, the task of protein function determination is a fundamental problem in biology. Thankfully, this problem yields itself to a computational solu-

tion, and a plethora of computational tools exist to supplement the experimental approaches. These tools range from sequence comparison to text-mining and are reviewed in more detail in Chapter 4

Likewise, we recognize that *in silico* annotation is an attractive solution, and present two novel approaches to solving this problem in Chapter 4. Briefly, to generate predictions, we integrate the sprawling Gene Ontology (GO) database, which represents the current ground truth of protein function annotation, into two end-to-end pipelines. In the first approach, we represent GO as a flat matrix and use Non-negative Matrix Factorization (NMF) to assign novel GO term annotations to proteins. In the second approach, we abstract Gene Ontology data by creating a protein-protein network, and use it as a conduit for function prediction via Global Information Diffusion, a method that propagates functional labels along edges in network. We provide cross-validation and retrospective validation for the two methods, and find that they are both capable of predicting novel annotations.

Chapter 4

Reasoning on Gene Ontology Networks Predicts Novel Protein Annotations

This work is under preparation for submission. Anticipated author order is Ilya Novikov, Angela Wilkins, and Olivier Lichtarge.

Author contribution: Conceived and designed experiments: IN, AW; Performed the experiments: IN; Analyzed the data: IN, AW. Writing the paper: IN, OL.

4.1 Introduction

In order to attenuate living systems, one needs to understand how individual components within the cell function and interact. As a result, much of the literature in biology is focused on modeling and validating function of individual proteins and nucleic acids. Experimental techniques that underpin function annotation in proteins such as protein microarrays [76], yeast two-hybrid screens [75], co-immunoprecipitation [77], and structure determination [78–80] are powerful but also expensive, and often scale poorly. As a result, the rate at which proteins are annotated is lagging behind the rate at which new proteins are discovered (Fig 4.1). Furthermore, because biomedical research tends to be narrowly focused, a minority of proteins account for the majority of annotations. For example, the upper third of the best-annotated human proteins in the Uniprot database [81] accounts for over 70% of all human annotations (Fig 4.2).

Because the cost of experimental annotation is high, the research community has produced a large number of computational solutions to the problem of functional annotation [82]. A number of function prediction methods infer function on the basis of sequence homology, making a resonable assumption that similar protein sequences share functional properties [83–85]. The same logic drives to structure similarity approaches that tranfer functional annotations between proteins based on similarity of their atomic structure [86–88]. A subset of these methodologies refine the homology aspect by focusing on local features like subdomains and structure motifs rather than the full-length molecule [89–91]. Another approach makes use of genomic features, hypothesizing that if two genes exist in a

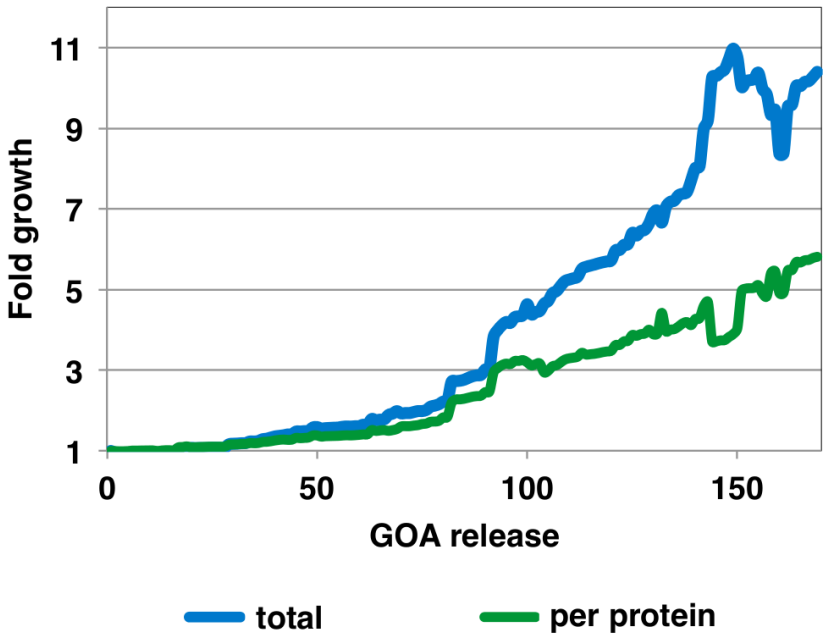

Figure 4.1: Annotation growth is outpaced by discovery of new proteins. Shown is the number of annotations with each successive GO annotations (GOA) release for UNIPROT database. While the total number of annotations, blue, is growing rapidly (due to growth in the number of new proteins), the average annotation density, green, is increasing at a much slower rate.

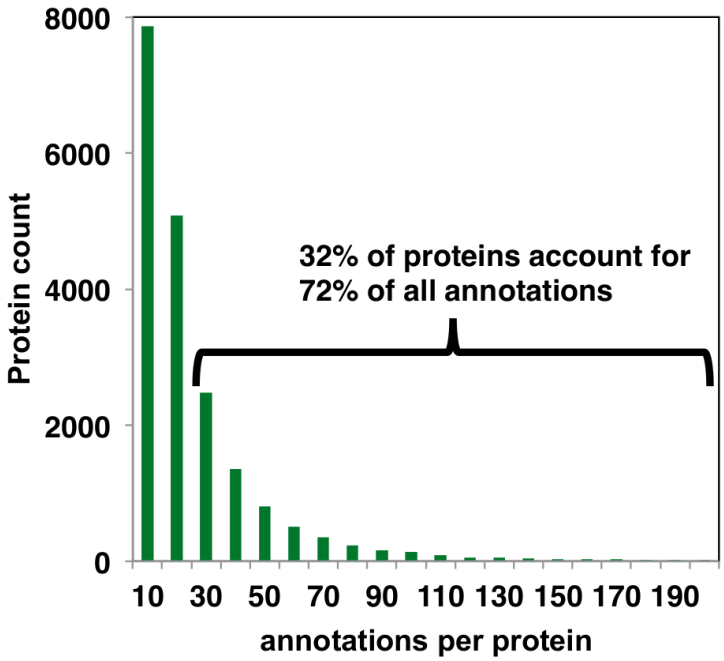

Figure 4.2: Well-annotated proteins account for most of the annotations. The plot is based on the June 2017 release of GO annotations for Uniprot.

similar genomic context, they are likely to share function [92]. Many methods make predictions on the basis of protein-protein association networks. Given a network, they assume that connected proteins are also functionally connected [52, 93, 94]. Finally, Natural Language Processing (NLP) algorithms infer functions by directly text-mining biomedical literature for evidence of functionality [95, 96].

Here, to supplement these existing approaches, we propose a novel computational solution that is based purely on the integration and analysis of the data contained in the Gene Ontology (GO) database [97]. Since its inception in 1998, GO has become the standard for protein classification in biology. GO organizes available evidence and classifies proteins by labeling them with GO terms. GO terms describe a proteins molecular function, biological process, and structural location within the cell. The terms are arranged in a hierarchy, the term ontology, according to their specificity and context. Higher-level terms describe general features, while radiating child terms provide increasingly granular descriptions (for example of a protein annotated with GO terms, see Table 4.1, and refer to Fig 4.3 for an example of a term hierarchy). The Gene Ontology has been so successful with its categorization schema, that the 2019 release of the database contains 597 million annotations, and over 46 thousand unique GO terms. Clearly, the Gene Ontology contains a trove of curated, context-sensitive biological data. We hypothesized that GO, because it is so information-rich, can be used as a sole feature set for predicting novel functional annotations.

Here we present two different models for extracting, modeling, and predicting on GO data. In our first approach (Fig 4.4), we extract all protein-term annotation

GO term id	GO term
GO:0002326	B cell lineage commitment
GO:0007569	cell aging
GO:0071479	cellular response to ionizing radiation
GO:0034644	cellular response to UV
GO:0007417	central nervous system development
GO:0051276	chromosome organization
GO:0008340	determination of adult lifespan
GO:0006302	double-strand break repair
GO:0048568	embryonic organ development
GO:0007369	gastrulation
GO:0001701	in utero embryonic development
GO:0071850	mitotic cell cycle arrest

Table 4.1: GO terms assigned to human protein P53. Note that the terms are not flat labels. Instead, they are arranged in a hierarchy that contextualize each term's identifier within a wider functional scope.

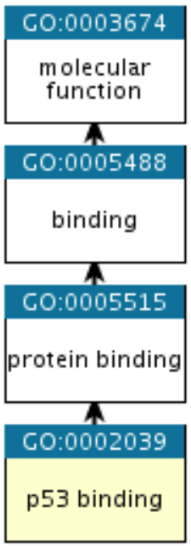

Figure 4.3: Ancestor-child term relationship. All GO terms are arranged into a heirarchy with lower-level terms providing a granular description of a general property or process designated by the ancestor. Shown here is an example of protein binding terms organized into a hierarchy.

for an organism from GO, and convert them into a sparse matrix A, where A(i,j) = 1 denotes that protein i is annotated with GO term j, while the remaining cells are set to 0. To predict novel protein-term pairs on this matrix, we employ Non-negative Matrix Factorization (NMF), a machine learning method used for matrix completion [98]. NMF factorizes the original matrix A, and then reconstructs it as an approximation A, where all of the possible protein-term pairs now have a weight assigned to them. The weights reflect the likelihood that a protein-term pair exists in the original matrix A, and therefore in nature.

Next, we developed an alternative model based on the concept of GO term semantic similarity [99], a measure that reflect how similar two proteins are according to their GO terms. To implement this model for an organism, we measure GO term similarity for every possible pair of proteins in an organism, and then withhold the lowest scoring edges from the matrix to create a sparse protein-protein network (Fig 4.5). In these networks proteins with similar GO terms preferentially connect to each other. To predict novel annotations on this network, we employ Global Information Diffusion [100] to propagate functional labels from annotation-rich nodes to nodes that lack annotation.

The individual components of these two frameworks have been used in biology before. NMF, famously used in business to match consumers with products, has been used in biology to predict protein-protein and protein-RNA associtations [101, 102], and to draw connections between proteins, drugs, and diseases [103]. Meanwhile, Global Information Diffusion has seen use in a variety of functional studies. In one example, GID was used to propagate P53 labels over

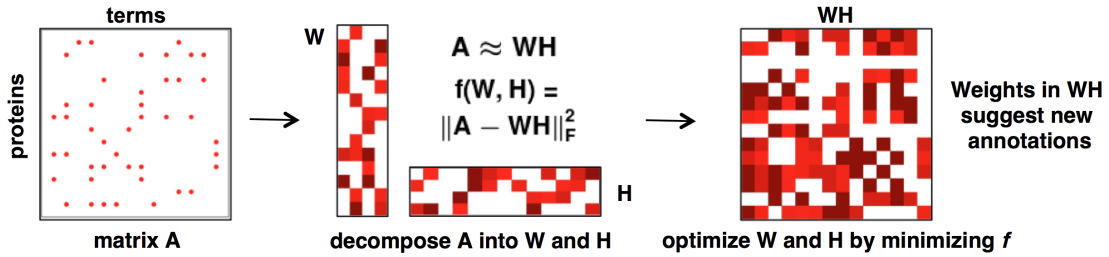

Figure 4.4: GO-NMF model for prediction of novel protein-term annotations. Annotation corpus for an organism is flattened as a matrix representation mapping proteins to their annotated terms. NMF factorizes this matrix by randomly seeding two smaller submatrices W and H, and then optimizing them, until the product WH approximates the original matrix. The imperfect approximation contains post-factorization scores that reflect how likely a protein is to be annotated with a term.

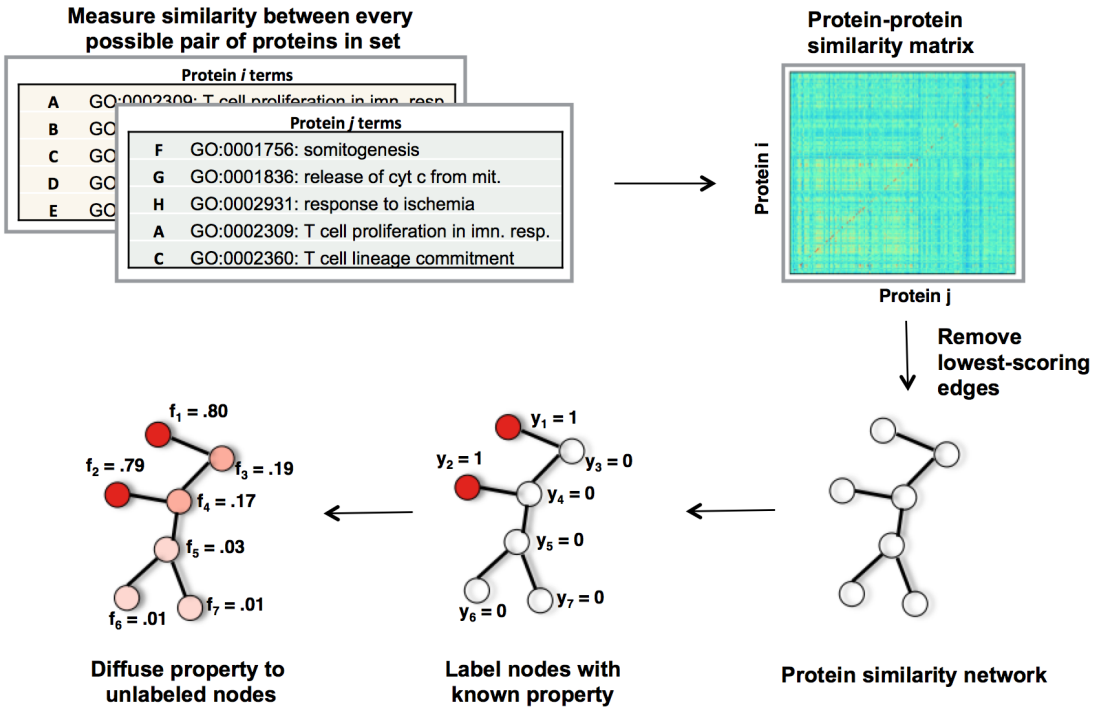

Figure 4.5: GO-GID model for prediction of novel protein annotations. For a set of proteins, we measure GO term semantic similarity in a all-vs-all manner to construct a similarity matrix, where every pair of proteins has a similarity weight associated with it. We convert the matrix into a network by dropping low-scoring weights. In the network, we can then label nodes with their known functions, and diffuse the labels to nodes with no annotations.

a literature-based kinase network to predict novel kinases of this important oncogenic protein [95,96]. In another example, GID over supergenomic protein network identified EXP1, a protein in *Plasmodium falciparum*, as a target for anti-malarial therapy. GID of functional annotations over a network of evolutionarily-important structural motifs had been used to assign functions to orphan proteins [52]. Lastly, GO term semantic similarity has been used to create networks to infer protein-protein interactions [104, 105]. These algorithms are robust and general, and integrating them with the Gene Ontology to create a predictive framework is novel.

We tested the two models, and found that both have predictive power. GO-NMF predicts future protein-term annotations in retrospective and time-stamped validation, and is general across species (species modeled shown in Table 4.2). To compare GO-NMF to other methods, we submitted predictions to the time-delayed Critical Assessment of Functional Annotation challenge [82], and found that NMF-GO performed on par with the field (publication in preparation by the CAFA Consortium). Meanwhile, we cross-validated GO-GID in the context of human kinase, human proteomic, and *P. putida* proteomic networks, and found that GO-GID recovers protein nodes with similar function, including nodes associated with specific phosphorylation targets, diseases, or GO terms. We then used GO-GID to help validate members of the PAK kinase family as interactors of oncogenic protein P53.

4.2 Methods

4.2.1 Gene Ontology

Before describing the predictive models, it is useful to first discuss the organization of data in the Gene Ontology. In practice, Gene Ontology consists of two sets of data. First is the annotation corpus, which is the collection of all annotations assigned in GO to the proteins of a particular species. GO maintains separate annotation corpora for popular model species such as humans and mice, as well as a massive aggregate depository that includes GO annotations displayed by the proteins in the Uniprot database. The second component of GO is the term ontology, which is the directed acyclic graph that describes the GO terms and their relationships. The ontology information is available as a single file and is general to all species. Note also that GO terms are organized are organized into 3 independent categories: molecular function ontology (MFO) terms to describe functional mechanisms, biological process ontology (BPO) terms to describe pathways, and cell component ontology (CCO) terms to describe the physical location. We downloaded the annotation corpora for the 19 test species (Table 4.2) and the ontology graph at `http://geneontology.org/docs/downloads/`.

4.2.2 Reasoning on GO with NMF (GO-NMF)

Representing Gene Ontology data for NMF

The first step in making predictions is to convert knowledge in the Gene Ontology into a mathematical representation amenable to reasoning. With Gene Ontology

Organism	Proteins	Annotations
Mycoplasma genitalium	438	3,733
Helicobacter pylori	1,131	8,242
Methanocaldococcus jannaschii	1,353	10,460
Sulfolobus solfataricus	2,050	12,503
Pseudomonas putida	4,068	26,377
Bacillus subtilis (strain 168)	3,405	28,952
Pseudomonas syringae pv. tomato	5,076	29,506
Salmonella typhimurium	3,691	32,116
Candida albicans	6,423	39,631
Escherichia coli (strain K12)	3,888	50,126
Dictyostelium discoideum	8,857	69,179
Schizosaccharomyces pombe	5,102	82,966
Saccharomyces cerevisiae	6,719	110,767
Drosophila melanogaster	12,560	152,127
Danio rerio	21,591	166,685
Arabidopsis thaliana	25,395	239,163
Rattus norvegicus	19,282	334,681
Mus musculus	21,484	559,780
Homo sapiens	19,255	566,176

Table 4.2: Species used in validation. We modeled predictions for all species here with NMF, and for human and *P. putida* proteins with GID. These species were selected for validation, because they were featured in the CAFA3 competition.

annotations, a natural representation is an adjacency matrix A that maps proteins against their GO terms. Formally, given a set of proteins $\mathbf{p}$ and a set of terms $\mathbf{t}$, $A(p_i, t_j) = 1$, if protein p_i is annotated with term t_j, and 0 otherwise. Note that because GO terms are arranged in a hierarchy, a protein annotated with a low-level child term also carries all of the child's ancestor terms. For example, a protein annotated with term *P53 kinase* is also implicitly a *protein kinase*. We explicitly include these relationships in A. The adjacency matrix provides a medium suitable for making predictions with NMF.

Non-negative Matrix Factorization Algorithm

To predict novel connections in the protein-term matrix, we use Non-negative Matrix Factorization (NMF). NMF decomposes the original matrix A, of size $m \times n$, into two smaller matrices W and H of size $k \times m$ and $k \times n$. W and H are basis vectors that compress the relationships found in A into a reduced number of dimensions $k << min(n, m)$. The goal of the method is to generate W and H such that their product, WH, approximates the original matrix, $A \approx WH$. This problem is solved by seeding a random W and H, and iteratively solving the following set of equations for H and W:

$$\begin{aligned} &\text{Solve} \quad W^T W H = W^T A \quad \text{for H} \\ &\text{Solve} \quad H H^T W^T = H A^T \quad \text{for W} \end{aligned} \tag{10}$$

until the objective function F converges to a minimum:

$$F(W, H) = ||A - WH||_F^2 \tag{11}$$

The dot produt WH decompresses the basis vectors and creates an imperfect approximation of A, where each cell now carries a non-zero weight $WH(p_i, t_j)$ that reflects how likely protein p_i is to be associated with GO term t_j. These weights are the basis for novel predictions. We normalize the weights by converting them into a z-score, and rank order all predicted protein-vs-term pairs in WH in descending order. We treat the highest-scoring pairs (p_i, t_j) as more likely to occur in nature, than pairs with lower scores. The GO-NMF framework is depicted in Fig 4.4.

4.2.3 Reasoning on GO with GID (GO-GID)

In addition to GO-NMF, we sought to create an alternative model by leveraging the concept of GO term semantic similarity. The goal of GO term semantic similarity is to measure similarity between two proteins by comparing their GO term annotations. High degree of similarity implies that a pair of proteins perform a similar molecular function, participate in the same biological process, and are located within the same cell compartment. In practice, GO similarity is a proxy measure for protein-protein association. This property of GO term similarity can be used to model protein-protein networks, where protein nodes are connected by edges that represent GO terms similarity of the two proteins. In this manner, we abstract GO into a condensed network representation, which is readily compatible with function prediction via Global Information Diffusion. The approach is

sketched in Fig 4.5.

Generating Protein-Protein Similarity Networks

To generate a protein-protein similarity network for a given set of proteins, we first measure similarity between each possible pair of proteins i and j within the set. This, in turn, is accomplished by measuring the similarity between individual GO terms that belong to a pair of proteins. A large number of term similarity measures are available [99], but their performance is relatively close [106], so we chose to use the Resnik method [107] for its simplicity of implementation.

Given two GO terms q and w, Resnik traverses the Gene Ontology hierarchy, and records all ancestor terms shared by q and w. Next, it measures Information Content (IC) of each ancestor, defined as:

$$IC = -log\frac{t}{N} \tag{12}$$

where t is the number of times the ancestor term has been used to annotate a protein, and N is the total number of protein-term annotations in the set. The IC produces large weights for low-level terms that are positioned deep in the hierarchy and used infrequently due to their high specificity. Likewise, high-level terms have lower IC weights because they are very general and are used frequently. Ancestor term with the highest Information Content is designated as the Most Informative Common Ancestor (MICA) of terms q and w. The Resnik similarity of two GO terms, then, is simply the IC weight of their $MICA$, formally:

$$similarity(q, w) = IC(MICA) \tag{13}$$

Given two proteins, then, we determine in a pairwise manner the similarity of their GO terms. To convert these term-term scores into a single measure of similarity for the protein pair, we employ Best Match Averaging (BMA). For each term, we keep the score of its most similar match, and average the scores across the set. This BMA score is the GO term similarity weight for the two proteins i and j, sim_{ij},. The higher the score, the more similar the two proteins are.

Finally, to create the protein-protein similarity network for a set of proteins of size N, we measure Resnik similarity sim for each possible pair of proteins i and j in the set, and record them in the similarity weight matrix W, so that $W(i, j) = sim_{ij}$. To convert this weight matrix into a graph, we iterate through each protein in W and set its $\sqrt{N}$ most similar partners to 1, and the remaining to 0. The result is a sparse protein-protein GO similarity network, G, where edges connect proteins that display a high degree of GO term similarity.

This network can now be used to propagate informationa across protein nodes via Global Information Diffusion.

Global Information Diffusion

The goal of Global Information Diffusion (GID) is to propagate information between neighboring nodes in the network. The utility of GID comes from the assumption that two protein connected in the network are also connected within the cell, and therefore functionally associated. By diffusing functional labels between

connected nodes, we can spread annotations to previously-unannotated nodes.

Information is initially seeded in the network by labeling proteins that display a relevant property with label $y = 1$, and setting the remaining nodes to $y = 0$. The label is then diffused across the network by minimizing the following objective function:

$$H = \sum_{i} (y_i - fi)^2 + \alpha \sum_{i,j} G_{ij}(f_i - f_j)^2 \tag{15}$$

where vector y contains the initial label information for each protein i, and vector f contains the protein's post-diffusion label weights. The first term in the function seeks to minimize the loss of the initial label (the term is minimized to 0 if the post-diffusion score f_i is the same as the initial label y_i). The second term counters this by maximizing label similarity between neighboring nodes i and j in the graph, G (note that the second term reaches 0 if the post-diffusion label of node i, f_i, is the same as the weight of its neighbor j, f_j). The balancing of these two terms propagates information through the network. (Diffusion factor α is used to control the contribution of the second term to H; larger α forces the function to maximize diffusion to neighboring nodes.)

The final output of diffusion, vector f, contains the post-diffusion label scores for each of the nodes, and higher scores correspond to a higher likelihood that node displays the labeled property in nature. For final application, we normalize weights in f as a z-score.

Evaluation Methodology

We use two standard tests to evaluate both models.

Cross-validation is the first test. The goal of cross-validation is to "hide" part of the existing data from the test set, and then attempt to recover it using the model. For example, given a protein-term matrix A, we divide A into 10 randomly-chosen partitions, or folds, and then set all non-zero edges in one of the partitions to 0 (we "hide" or "leave" that fold out). We then apply NMF to the remaining 9 folds to reconstruct the partition that has been hidden. The purpose of the "hidden" fold is to serve as a validation set, while the remaining folds are used to train the model. If the method is robust, it will assign high weights to the "hidden" edges. We repeat this for each of the 10 folds, until we cover every edge in A. We then evaluate whether resulting predicted weights agree with the original matrix A. A special type of cross-validation, when the number of folds is equal to the number of items in the data set, is known as leave-one-out validation (that is, each fold is of size 1).

The second test is retrospective (time-stamped) validation. Unlike cross-validation that assays the ability to recover pre-existing structures, the goal of retrospective validation is to simulate *de novo* predictions. We achieve this by splitting the input data into a training and a validation set according to a particular time stamp. For example, we exclude all annotations made after 2015, and then use the available data to make predictions. A robust predictor will produce predictions that closely agree with actual post-2015 annotations.

We evaluate predictor performance in both of these experiments using Receiver

Operating Characteristic Curve. We discussed ROC in Chapter 2, but let us briefly review. A robust prediction algorithm will assign, within its particular scoring schema, a higher score to a true interaction (in our case, a higher factorized score in WH, or a higher post-diffusion label y), than to an interaction that does not exist. Given a set of rank ordered predictions, we traverse them in order of decreasing score, and at each unique threshold, evaluate both TPR and FPR. Plotted as curve, TPR will increase at a faster rate than FPR if the predictor provides robust classification. An ideal classifier (with TPR = 1, and FPR = 0) will produce a ROC curve with AUC=1. A predictor that lacks discriminatory power will produce a curve with AUC=0.5. We use the area under the ROC curve (AUC or AU(RO)C) as a primary evaluation metric in this chapter.

4.3 Results and Discussion

4.3.1 GO-NMF Predicts Novel Annotations

To determine whether GO-NMF can be used as a predictive tool, we first cross-validated NMF in a set of three model species, *B. subtillis*, *S. cerevisiae*, and *D. Melanogaster*. For each of the species, we downloaded the annotation corpus from GO, and converted the annotations into an adjacency matrix of protein vs terms. All existing annotations were marked with 1, and the rest were left unassigned (set to 0). Next, we tested NMF's prediction capacity to predict novel edges in these matrices via 10-fold cross-validation. Briefly, we divide the adjacency matrix it into 10 folds and "hide" one of the folds from the data set. Then, we apply

NMF to the remaining data to recover the left-out fold. We iteratively repeat the experiment for each fold, and compare post-factorization edge weights to the original matrix. Data are shown in Fig 4.6 as ROC curves.

NMF performed well in all 3 species, with AUC=0.97 in *B. subtillis*, AUC=0.98 in *S. cerevisiae*, and AUC=0.99 in *D. Melanogaster*. To test whether this performance is general, we applied the cross-validation experiment to 11 additional species. We found that NMF performed equally well in these species, and average AUC across all test cases is 0.95 (Fig 4.7). These data show that NMF assigns high post-factorization scores to the "hidden" protein-term edges during cross-validation, and that this predictive power is general across multiple species.

Next, to better gauge GO-NMF as predictor of *de novo* connections, we subjected NMF to restrospective validation. We hid annotations discovered after January 2016 from the data set, and generated the protein-term matrices for the three model species. We then applied NMF to these time-stamped matrices to generate predictions. We compared the predictions against protein edges actually discovered after January 2016, and found that NMF recovers them with high AUCs: AUC is 0.88 in *B. subtillis*, 0.91 in *S. cerevisiae*, and 0.89 in *D. Melanogaster* (see Fig 4.8). To test whether this performance is general, we expanded the retrospective analysis to 16 additional species (Fig 4.9), and found that mean AUC across all test cases is 0.89. These data show that NMF-GO is able to suggest novel protein-term edges before they appear in the Gene Ontology.

We next asked, how can the GO-NMF model be improved? One simple avenue for improvement is to increase the size of the input data set by adding more

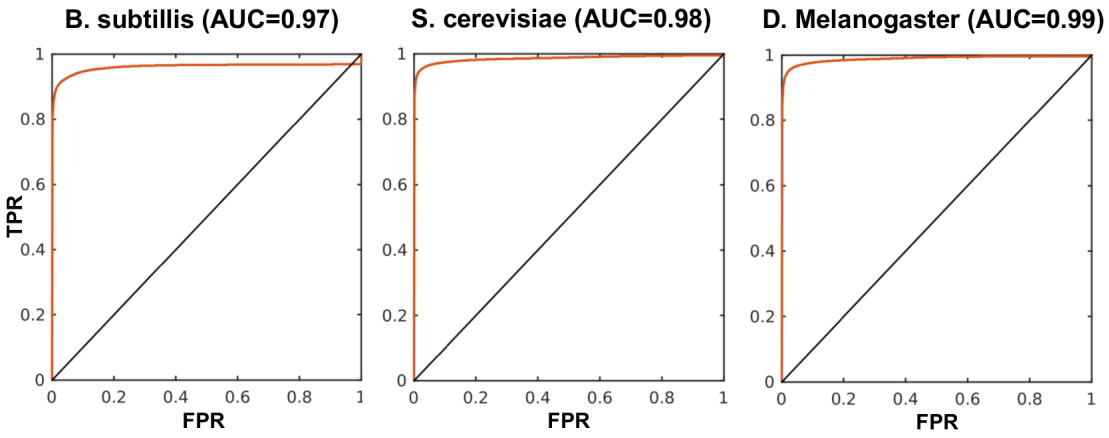

Figure 4.6: Cross-validation of GO-NMF in three model species shows that NMF has predictive power. In all three species, AUC is significantly greater than 0.5. Each curve represents a cross-validation experiment, whereby the input data are divided into 10 folds, followed by an iterative process of leaving one fold out, and using the remaining 9 to generate predictions. Once predictions are produced for each of the left-out folds, the reconstituted matrix is compared to the original input matrix of protein-terms.

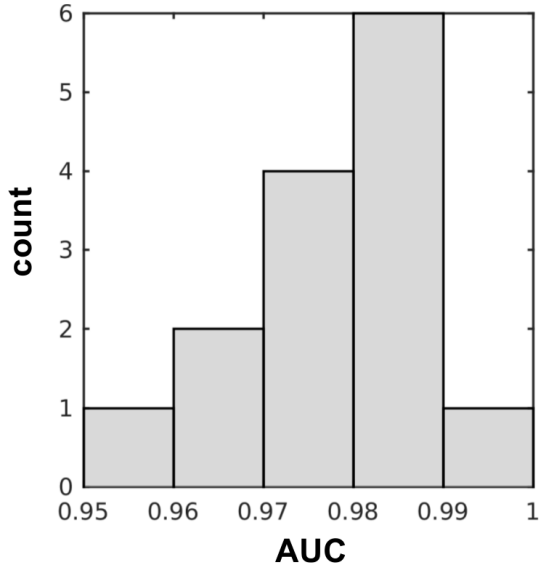

Figure 4.7: NMF predictive power is general across species. We extend the cross-validation experiment to 11 additional species (14 total), and find that in these species NMF performs equally well. Mean AUC across all 14 species is 0.95.

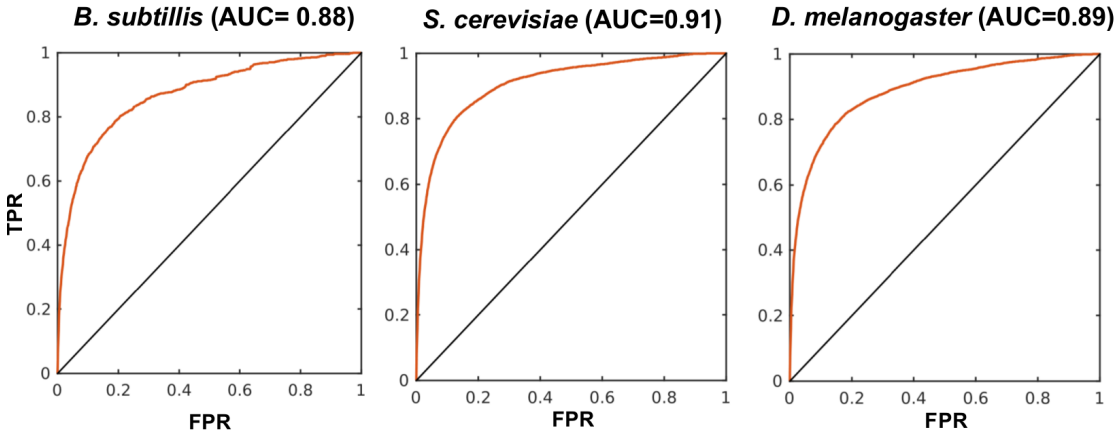

Figure 4.8: Time-stamp validation of GO-NMF in the three model species. For each of the 3 species, we generated a matrix of input data based on protein annotations found in the January 2016 release of GO. We then applied NMF to these data to make predictions, and found that the predictions are enriched for edges (annotations) discovered after January 2016. This further shows that GO-NMF has predictive power.

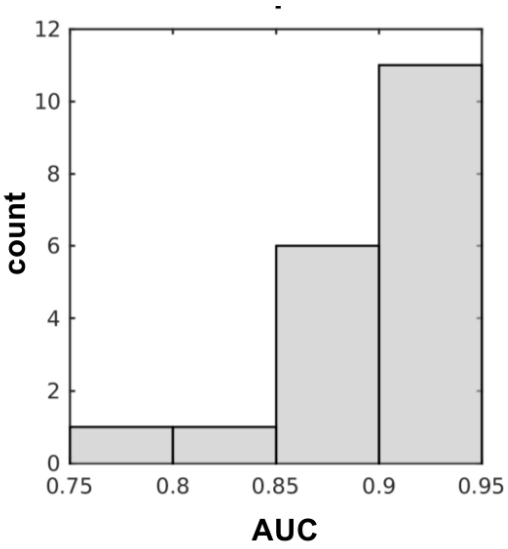

Figure 4.9: GO-NMF predictive power is general across species. We extended the time-stamped validation analysis to 16 additional species (19 total), and found that NMF predicts time-stamped edges with a mean AUC of 0.88.

features. We tested this hypothesis in two experiments.

For three test species, we segregated annotations into three independent input matrices A_{BPO}, A_{MFO}, A_{CCO} according to the ontology of the annotation (biological process ontology terms are delegated to A_{BPO}, and so on). We then conducted 10-fold cross-validation on each of the individual ontology matrices. We found that NMF's performance on individual subontolgoies is robust, but is lower than in the full ontology. In the three examples in (Fig 4.10), the average performance of the entire ontology was AUC=0.976, and AUC=0.958 for the individual ontologies. These data suggest that combining all annotations into a large inclusive dataset, as we have done from the start, yields better predictions.

In another experiment, we examined retrospective performance of NMF in *D. discoideum*. The number of annotations in *D. discoideum* steadily increased from approximately 24,000 in 2012 to 44,000 in 2017. We hypothesized that the growing number of annotations should lead to better performance of NMF in 2017, compared to 2012. To test this hypothesis, we conducted a series of retrospective experiments. Starting with *D. discoideum* annotations available in 2012, we attempted to predict new annotations that had appeared by year 2017. We then moved the time stamp closer to 2017 one year at a time, and repeated the experiment. We discovered that as the input annotation set grows and approaches 2017, the NMF recovers novel annotations with higher accuracy (Fig 4.11). These data are in agreement with the ontology segregation experiment; greater number of input features leads to better predictions. These data also suggest that performance of NMF will continue to improve in the future, simply as a byproduct of

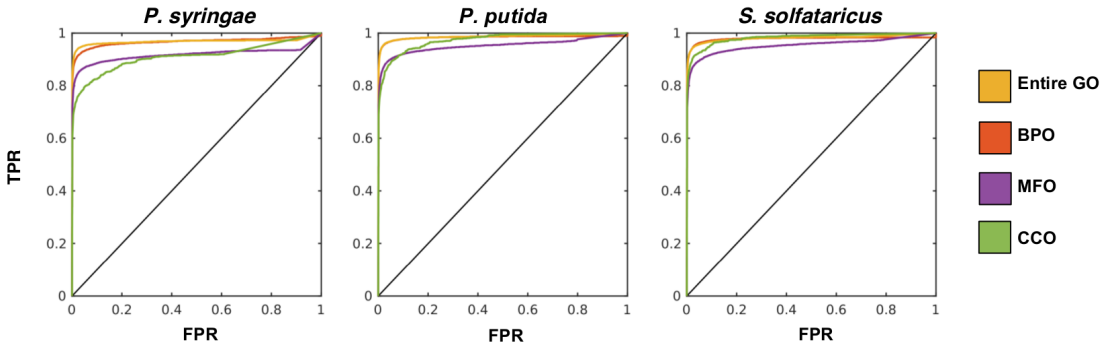

Figure 4.10: Combined ontology matrix improves performance. Predictions generated by the combination of the three ontologies perform slightly better than individual ontologies. The Cell Component (CC) ontology, which is not as densely populated as the Biological Process Ontology (BPO) and the Molecular Function ontology (MFO), has the lowest predictive power.

Gene Ontology growth.

Finally, as a test of prospective validation, we submitted our prediction for the 19 species to the Critical Assessment of Function Annotation (CAFA) competition [82]. The goal of CAFA is to survey the field of function prediction by soliciting predictions from the community. Participants are allowed 6 months to make and submit predictions; once predictions are submitted, the organizers wait 6 months to allow the Gene Ontology to accumulate new annotations. Then, they evaluate predictions made by the participants against these novel annotations, and rank the competing methods. We submitted protein-to-term term predictions made with GO-NMF, and found, according to the preliminary ranking released by the consortium, that, averaged across all prediction categories, our method ranked 36th out of 62 participants. GO-NMF performed better in predicting BPO go terms (ranked 15th), as opposed to MFO (39th) and CCO terms (45th), which is expected as BPO is the denser of three ontologies. A more detailed evaluation of these data is forthcoming in the consortium publication. However, the rankings suggest that GO-NMF, a fairly basic approach, performs on par with the field.

In summary, NMF-GO provides a simple and intuitive schema for predicting novel term-to-protein annotations in the Gene Ontology. The method performs well in both cross-validation and time-stamped validation, and produces better predictions when operating on a larger data set. Finally according to CAFA3 results, GO-NMF performs on par with other methods in the field, which is encouraging for such a simple paradigm.

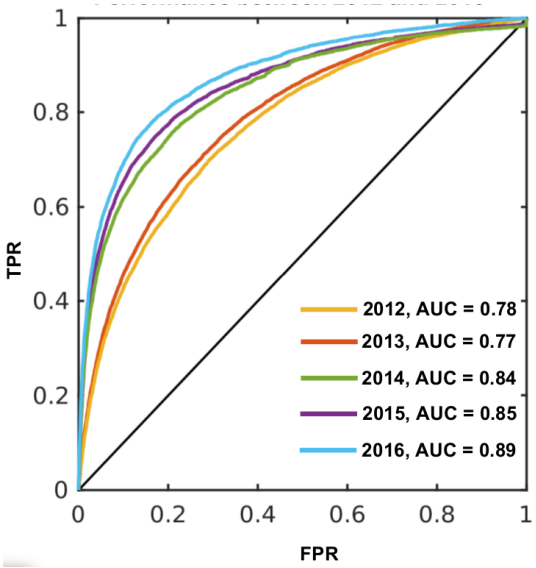

Figure 4.11: Retrospective performance in *D. discoideum* improves as more data becomes available. From 2012 to 2016, the number of annotations in *D. discoideum* nearly doubled, and that growth is reflected in better time-stamped performance through the years.

4.3.2 GO-GID Predicts Novel Annotations

While the data show that NMF-GO is a viable approach for reasoning on GO, the NMF is a flat, minimalist representation of GO. It includes relationships between proteins and terms, but it does not encode the term-term relationships contained in the GO hierarchy. To explicitly incorporate term-term relationships into our predictions, we designed an alternative reasoning schema for GO, rooted in the concept of GO term semantic similarity. GO term semantic similarity is a measure that reflects the similarity between any two proteins, based on the comparison of their GO terms. Given a set of proteins, we can measure in a pairwise manner the degree of similarity between each of them. This produces a protein-protein matrix, M where each weight $M(i, j)$ represents an edge connecting two proteins i and j. We convert this weight matrix into a graph (network) by explicitly removing all of the lowest-scoring connections (setting $M(i, j) = 0$). The remaining pairs in M are set to 1 to represent an edge connecting the two proteins. In this manner, we produce a protein-protein network.

This network representation of GO is readily amenable to reasoning by Global Information Diffusion. Briefly, the goal of GID is to diffuse information along the edges of the network, and the basic hypothesis is that two proteins closely connected in the network are more likely to share traits, such as function. To propagate information, we label known proteins with a label of interest (set corresponding label in vector y to 1), diffuse the label over the network to other protein nodes (initial label $y = 0$), and rank proteins according to their post-diffusion score f. Higher post-diffusion scores indicate a higher likelihood for the protein node

to display the trait designated by the label. In this manner, we suggest novel functions for previously unannotated proteins. For more detail on GO semantic similarity and GID see Methods.

We first tested this model on a GO similarity network of 509 human kinases. To construct the network, we compiled the list of all known human kinases (n=509), and for each kinase collected its associated GO terms from the human annotation corpus. Next, we measured GO term similarity between each possible kinase pair, and created the GO similarity weight matrix. For each protein in this matrix, we kept 25 highest-scoring edges, and set the remaining to 0. This produced in a human kinase network, where each node is connected to its most similar kinase by an edge. Note that because Resnik semantic similarity only measures similarity between pairs of GO terms from the same ontology, we created three separate networks, each based on the Biological Process, Molecular Function, or Cell Component annotations.

To test whether these three networks have predictive power, we examined kinases that phosphorylate P53, a major oncogene. We compiled a list of known P53 kinases (Table 4.3), and labeled their nodes in the network. We then subjected these labeled nodes to leave-one-out cross-validation. As described before, the goal of cross-validation is to "hide" part of the data, and then test whether the predictor recovers it. Here, we "hide" P53 kinases, one at a time, by setting their node's functional label (for association with P53) to 0. We then diffuse the remaining P53 labels via GID, and record the post-diffusion f-score for the left-out kinase. We iteratively repeat this until we have acquired an $f - score$ for each P53 kinase

in positive set.

If the prediction method is robust, then the left-out P53 kinases will be ranked highly compared to other, non-P53 nodes (the negative set). The data, in Fig 4.12, show that to be the case. GID recovers left-out P53 kinases with AUC=0.87 in the Biological Process-derived network, AUC=0.87 in the MF network, and AUC=0.69 in the CC network. These data show that in the GO similarity kinase network P53 kinases are connected, and are predictive of each other.

One interesting finding from cross-validation is the disparity in performance between the three ontologies. The BPO- and MF-derived network clearly performs better than CC. This is not surprising as CCO is the shallowest of the three ontologies, and is in line with our cross-validation experiment in GO-NMF that showed that CCO ontology lags in performance (Fig 4.10).

Another interesting finding is the impact of automatically-assigned annotations on performance. Roughly half of all annotations in the Gene Ontology are assigned automatically without human curation (IEA code in the annotation corpus). It is a reasonable assumption that IEA annotations aren't as reliable, and therefore would negatively impact prediction. To test this hypothesis, we excluded IEA annotations from the input data, and constructed the networks based solely on manually-curated annotations. The data (*NOIEA* curves in Fig 4.12) show that IEA-deprived networks performed significantly worse across all three ontologies. These data suggest that IEA anotations are actually quite reliable, and that the method benefits by including all of the available annotations, much like GO-NMF.

So far the validation has been focused on P53 kinases, so next we asked whether

P53 Kinase	Year discovered
CSNK2A1	1990
CDK1	1992
PRKDC	1992
CDK2	1995
MAPK8	1995
CDK7	1997
CSNK1D	1997
MAPK9	1997
EIF2AK2	1999
CHEK1	2000
CHEK2	2000
GSK3B	2001
MAPK1	2001
PLK3	2001
AURKA	2004
TAF1	2004
RPS6KA3	2005
CDK9	2006
CDK5	2007
DYRK2	2007
HIPK2	2007
IKBKB	2007
TTK	2009
AURKB	2011
CSNK1A1	2011
RPS6KA1	2011

Table 4.3: P53 kinases and their date of discovery. The list had been provided by A. Wilkins.

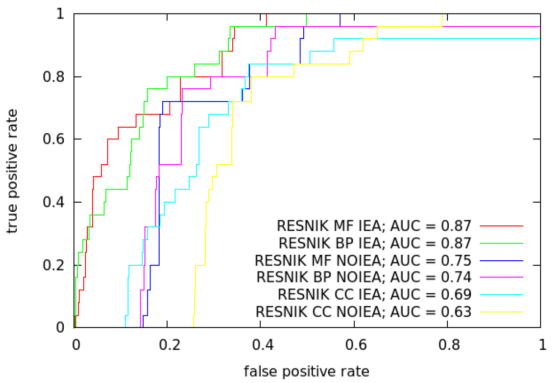

Figure 4.12: P53 kinases Leave-One-Out validation. These data show that GID recovers P53 kinases at a high rate, and that the underlying network is robust. Additionally, these data show that CCO ontology is not as predictive as BPO and MFO, and that excluding automatically-assigned annotations degrades peformance ($NOIEA$ in the legend denote networks built without IEA annotations). Note also that these networks were constructed without the P53-related terms such as the *P53 binding* term, in order to avoid trivializing the problem.

the network accurately represents links between other kinases. We compiled a list of 41 different phosphorylated proteins, and for each protein identified the list of its kinases (from the PhosphoSite database [108]). Just like with P53 kinases, we used the phosphorylated protein-kinase association as the functional label of interest. We subjected these 41 kinase sets to LOO cross-validation in the BPO network, and found that they perform as well as the P53 kinases. The data is summarized in Fig 4.13. 88% of the kinase sets performed with an AUC greater than 0.8. These data show that the kinase network reflects connections between all kinases not only the kinases that phosphorylate P53.

To further test the generality of the model, we created a human proteomic network using the biological process annotations from the human annotation corpus. In this network, we tested for links between proteins according to their disease association. We compiled a list of 1571 human diseases, and to each disease assigned the proteins associated with it in the OMIM database [109]. We applied LOO validation to each of these disease-associated protein sets. We found, in Fig 4.14, that GID recovers disease genes with AUC > 0.70 in 84% of the cases. These data show that disease-associated genes cluster in the network of human proteins, and are predictive of each other.

For our last test, we expanded the cross-validation into a non-human system. We constructed a proteomic-scale BPO network for *P. putida*, and in this case used Gene Ontology terms as our functional label of choice. We went through each term one at a time, and labeled all of its associated proteins in the network, then conduct LOO validation to see if proteins annotated with the term would

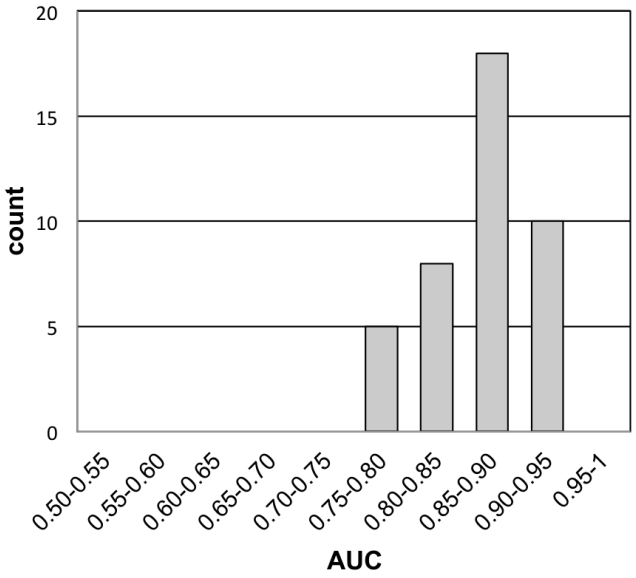

Figure 4.13: Performance of 41 kinase sets in LOO cross-validation shows that the kinase network is general to non-P53 kinases. The kinases are recruited into sets based on their target protein, each set is then subjected to LOO validation.

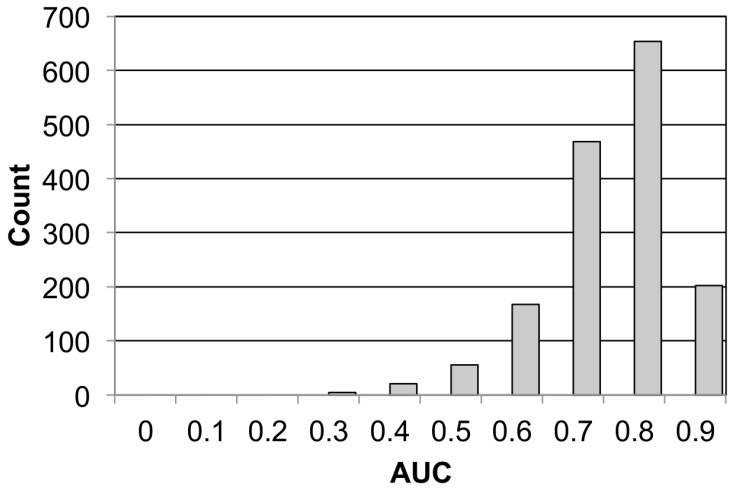

Figure 4.14: GID recovers disease-associated proteins in the human proteomic network. Each point in the histogram represents a single LOO experiment conducted on a set of proteins associated with a disease, and disease association is the functional label being predicted during LOO validation. Disease-protein relationships were extracted from OMIM by B. Bachman.

rescue their "left-out" sibling. The data, summarized for all sets as a single curve in Fig 4.15, shows that GID recovers "hidden" term labels with AUC=0.87. These data show that to modeling GO is general across species.

Together, these data confirm our primary hypothesis: abstracting Gene Ontology via term similarity produces a network with node connections that reflect protein-protein associations occurring in the cell. This gives the network predictive power.

Prospective Predictions and Validation

Retrospective studies have shown that GO similarity networks are predictive when combined with GID.

Next, we tested whether the approach can be used to predict future annotations in a series of time-stamped trials. First, we tested on the P53 kinases. We constructed a series of kinase networks spanning year 2003 to 2011, and in each of the networks labeled kinases known at that time to phosphorylate P53. We then diffused them to predict annotations, and tested the prediction results against the P53 kinases which had not been discovered yet. In this manner, we assayed networks between 2003 and 2011, and the data are plotted in Fig 4.16. The data show that there is modest positive enrichment for yet-undiscovered P53 kinases, and that performance improves over time. Next, as a test for generality, we also applied the retrospective test to the *P. putida* network. We created a network based on July 2016 annotations, and then diffused existing GO terms, in order to predict GO terms annotated to *P. putida* in the following the next 6 months.

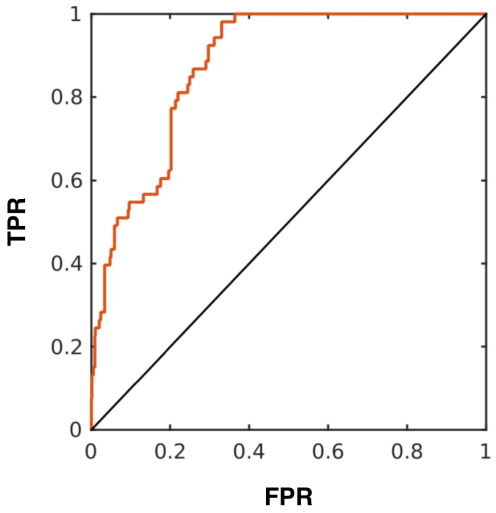

Figure 4.15: *L. putida* LOO validation is in agreement with other models. Here, we built the network based on the BPO, and diffused GO term labels. The performance is in line with our observations in the kinase network and the human proteomic network.

We found that diffusion predicted the new annotations with AUC of 0.71 (Fig 4.17). Together these data show that GO has some capacity to predict future annotations.

Finally, we compared our predictions to the novel, experimentally-validated P53 kinases discovered in the lab by L. Manley. She hypothesized that the family of p21(CDC42/Rac)-activated kinases (PAKs) regulates P53. The line of inquiry was supported by literature, but we also validated it computationally. We diffused the known P53 kinases on an up-to-date kinase network, and found that PAKs ranked in the top half of all predictions (see Table 4.4), including PAK2, PAK4, and PAK7 that ranked within the top 20%. When L. Manley tested the PAKs experimentally, she found that all phosphorylate P53, and two, PAK6 and PAK4, form stable interactions with it in vivo.

In addition to PAKs, A. Wilkins in the lab had predicted kinase NEK2 to be a regulator of P53, using P53 label diffusion on a literature-derived network of human kinases [96]. The lab of L. Donehower at Baylor College of Medicine tested the prediction experimentally, and confirmed that NEK2 is a negative regulator of P53. Diffusion of P53 kinases on the GO term similarity network places NEK2 in the top 10% of all predictions. Position of NEK2 and PAKs along ROC curve is show in Fig 4.18. These data show that GO-GID provides moderately strong support for the interaction of these proteins with P53, which agrees with the experimental data.

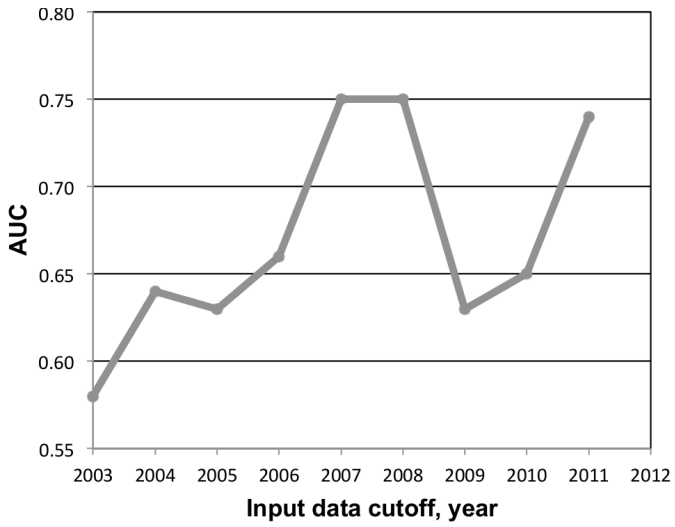

Figure 4.16: P53 kinases retrospective validation on networks with a time-stamp cutoff of 2003 to 2011. Each experiment attempts to recover future P53 kinases by diffusion of P53 labels known to be true in that year, using networks which are likewise constructed only on the data available up to that date. This validation set up is meant to imitate *de novo* prediction workflow. The data suggest there is some enrichment for P53 kinases among top-scoring prediction. Additionally, performance improves through the years, likely a result Gene Ontology's growth, as a greater number annotations creates a more reliable network.

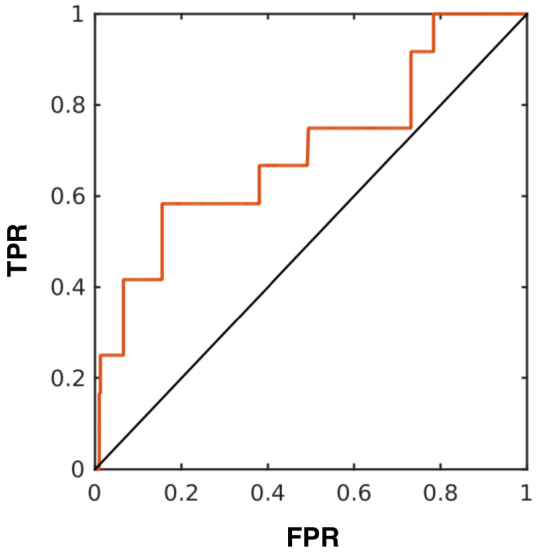

Figure 4.17: *L. putida* time-stamp validation. Here, we build a proteomic network for *L. putida* using annotations available in July 2016. We attempt to recover GO terms, assigned to *L. putida* in the following 6 months, by diffusing GO terms known in 2016. With AUC significantly greater than 0.5, These data show that these future annotations are ranked higher than average by the GO-GID predictor.

Node	label z-score	rank (out of 484)	% rank
NEK2	1.659	47	9.7%
PAK2	1.360	76	15.7%
PAK4	0.972	93	19.2%
PAK7	0.959	95	19.6%
PAK6	0.092	145	30.0%
PAK3	0.000	178	36.8%
PAK1	-0.071	210	43.4%

Table 4.4: GID ranks of PAKs and NEK2. GID ranks PAKs and NEK-2 in the upper half of the predictions for P53 interactions. Average percentile rank of these nodes is 25%.

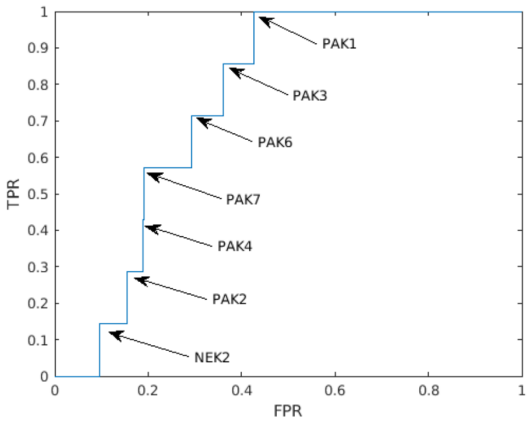

Figure 4.18: PAK and NEK2 ranks along the ROC curve. This is a visual representation of data in Table 4.4. Compared to other nodes in the network, GID properly identifies these kinases as more likely to interact with P53.

4.4 Conclusions and Future Directions

In this work, we set out to determine whether the Gene Ontology can be used as a novel predictor of protein function. We developed and validated two different approaches. The first, GO-NMF, applies NMF over an adjacency matrix of GO terms and proteins to predict new protein-term pairs. GO-NMF recovers existing pairs in cross-validation, and predicts future pairs in time-stamped studies. Performance of GO-NMF in CAFA3, while not spectacular, suggests that the method provides signal for genuine prospective prediction. The second approach, GO-GID, abstracts protein Gene Ontology terms to produce a protein-protein network, and reasons on it using Global Information Diffusion. Likewise, the method performs reasonably well in cross- and time-stamp validation, and is generally applicable in different contexts, such as predicting kinase targets in the kinase network, or predicting disease-associated genes in the human proteome. GO-GID predictions for P53 association for PAKs and NEK2 are in agreement with experimental data, though the ranking is not as sensitive as we would like.

The limitations of the method are informed by its basic implementation. The method relies squarely on a protein having meaningful annotations in the Gene Ontology in order to make predictions. A protein that is completely orphaned, or is only annotated with a handful of high-level terms, does not have enough features to create useful basis factors in NMF. Likewise a poorly annotated protein will be dropped from the similarity network at the thresholding stage, because it has no high-similarity edges connecting it to the rest of the network. Another limitation of the GO-based method is the lag between the initial discovery of a functional

feature, and its incorporation into the Gene Ontology by the human curator. That means GO-based networks will be slightly behind the current ground truth in literature.

The future direction is two-fold. First, the two models provide a lot of latitude for tuning of hyperparameters. In the NMF model, for example, we deployed basis factor $k = 200$, which is in line with other use case of NMF on proteomic-scale data. However, tuning k for each species to achieve the best granularity during factorization could improve performance. Likewise, in this work, we ignored the fact that decomposition of the matrix in NMF is seeded randomly, so every time the minimization function is solved, the solution is unique. To increase consistency of predictions, we could run multiple decompositions of the same matrix, followed by averaging of the post-compression weights. GID can be similarly tuned. For example, edges in the network graph G do not have to be binary. Rather than threshold, we can simply leave the continuous weights in place, providing more detail to the protein's local topology. Finally, another avenue for improvement is to combine the two models outright: NMF can be applied to the protein-protein network to predict novel edges, while GO-GID can be used to suggest likely protein-term pairs, which can then be included in the adjacency matrix for NMF.

Second, an attractive use case scenario for GO-NMF and GO-GID is to support other approaches. For example, the GO-NMF protein-term matrix can be directly combined with other matrices, such as protein-protein, and protein-drug matrices. Likewise, because GO terms are generally applied not only to proteins, but also to diseases, RNAs, and chemicals, GO-GID too can be used to create and reason

on combined protein-RNA-disease-drug networks. Futhermore, predictions made with GO-NMF and GO-GID can be readily combined with rankings produced by third-party methods. We briefly tested this by combining GO-GID predictions from the kinase cross-validation study with results of the identical experiment conducted on the literature-based Bag-of-Words network used in the NEK2 study [96]. Individual performance of the two individual networks is similar (mean AUC=0.84 for BOW, and 0.85 for GO). However, combining their $f - scores$ into a single output via logistic regression improves performance to AUC=0.89.

To conclude, we have abstracted data in the Gene Ontology into representations suitable for reasoning, and combined them in a novel manner with NMF and GID. The product is a complete end-to-end prediction pipeline based entirely on the Gene Ontology. Both of the proposed models have predictive power, and can be readily integrated with other methods.

Chapter 5

Conclusion

In this thesis, we have addressed the problem of function prediction in biology in two different contexts.

In the context of nucleotide analysis, we adapted a powerful sequence analysis tool, the Evolutionary Trace, to accept nucleic acid sequences. We provided a thorough validation of the method in the hammerhead and the bacterial ribosome, and found that ET accurately identifies functional sites in these molecules. We then expanded the analysis to families in Rfam, and found that major hallmarks of ET, the clustering of ET nucleotides and their overlap with functional sites, are a feature general in diverse RNA families. Finally, we have suggested two alignment optimization strategies that improve ET performance in RNA.

The study highlights that evolutionary principals which made ET successful in proteins, translate directly to structured fRNA, and suggest ET can be successfully be applied in other non-proteins contexts like DNA and viruses. As the scientific community's awareness of non-coding elements and their role in human diseases continues to grow, the need for a sequence analysis tool to help guide research will become more pronounced. RNA ET helps advance the field of non-coding RNA and DNA research by addressing this need.

RNA ET does have two significant limitations. First, the main output of ET analysis, ET rank of a nucleotide, is a fairly non-descriptive measure. While ET ranks nucleotides in relation to each other, it does not provide any contextual information on its own. ET does not show whether a nucleotide is strucutrally or catalytically-important, whether it is involved in a non-canonical pairing, or whether it is part of a structural motif such as a pseudoknot, a sarcin-ricin loop, or

a g-bulge (for review of RNA strucutral motifs see [33]. ET provides a hypothesis in form of predicted evolutionary importance, but it is up to the user to contextualize ET ranks within their molecule of interest.

One way to make ET predictions more nuanced is to incorporate nucleotide co-evolution into ET ranking schema. Co-evolution has a long history of application in RNA structure modeling [39–41], where it has been used to discriminate between paired and unpaired bases. The basic premise is straight-forward: given two nucleotides that form a Watson-Crick (WC) pair in the structure, a mutation in one nucleotide should be followed by a compensatory mutation in the other if the structural contact is to be maintained. Thus, the two paired nucleotides co-evolve, and the history of their co-evolution can be quantified using formal metrics such as Mutual Information [110].

Implementing co-evolution scoring for RNA would give us a second measure by which to discriminate bases. For example, two ET nucleotides that display a high degree of co-evolution are most likely paired in the structure. Examining the identity of the bases would further tell us whether the two bases are part of a canonical WC pair, likely to be found in a double-stranded stem, or a non-canonical interaction such as between bases in two kissing loops. Likewise, high-ranked ET nucleotides that does not display strong co-evolution could be assumed to be unpaired, and located in bulges or single-stranded loops. Incorporating co-evolution into RNA ET would not be techinically challenging, as this has already been done for protein ET in form of pair-interaction (pi)ET [56].

The second limitation of ET is its requirement for homologous sequences. Little

can be done to circumvent this requirement, as Evolutionary Trace is essentially a survey of sequence changes encoded in homologue divergence. This requirement precludes analysis of the most nascent RNA families (notably, lncRNAs), because their evolutionary history is too shallow. Here, researchers would be best served by the traditional tools of RNA analysis, such as structure prediction based on free-energy minimization or chemical probing [111].

In the second part of this work, we contribute to the field of protein function prediction. Rather than model sequences or structures to predict function, we addressed the problem from a Big Data perspective. We recognized that a key biological database, the Gene Ontology, has become so information rich, that directly reasoning on it to predict novel discoveries became a viable option. The well-ordered nature of the Gene Ontology and the human expertise encoded in the term hierarchy made GO even more attractive for prediction. To test this proposition, we designed two novel models for data extraction and reasoning on GO, and showed that both models have predictive power.

While being a useful display of GO utility beyond simple categorization, the overarching significance of the work lies in addressing the fundamental problem of function annotation in biology. Because there are tens of thousands of proteins and a similarly high number of possible functions, it is practically unfeasible to test all possible protein-function pairs experimentally. GO-NMF and GO-GID provide a simple and scalable approach for interrogating protein-term and protein-function associations computationally. By suggesting likely protein-function associations, GO-NMF and GO-GID can help guide experimental inquiry in protein research.

The main limitation of both models presented here, not surprisingly, is their reliance on pre-existing GO annotations. A novel protein that is completely unannotated cannot be incorporated into either model, because it has no GO terms. A protein that is sparsely annotated suffers from the same problem. The few high-level terms do not provide enough specificity to generate meaningful predictions about low-level functions. Proteins best suited for GO-based prediction are reasonably well-annotated proteins, for which our models can "fill in the gaps" by suggesting functions not obvious from manual assessment. However, for the entirely novel proteins, structural or sequence homology analysis remains the most tenable way to infer likely function. This conclusion is supported by the somewhat unimpressive performance of GO-derived models in the CAFA challenge, though it remains to be seen which methods scored best. Accordingly, we see the application of our GO-derived models primarily as a source of additional input features into a meta approach that incorporates multiple analyses into a single prediction. In this context, our models are valuable because they provide a laconic encapsulation of the information presented in the Gene Ontology.

Bibliography

[1] F. H. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–63, 1958.

[2] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970.

[3] Niles Lehman. RNA in evolution, journal = Wiley Interdisciplinary Reviews: RNA, volume = 1, number = 2, pages = 202-213, abstract = RNA has played a variety of roles in the evolutionary history of life on the Earth. While this molecule was once considered a poor cousin of the more influential polymers in the cell, namely DNA and proteins, a string of important discoveries over the last 50 years has revealed that RNA may in fact be the cornerstone of biological function. In particular, the finding that RNA can be catalytic, and thus possess both a genotype and a phenotype, has forced us to consider the possibility that life's origins began with RNA, and that the subsequent diversification of life is aptly described as a string of innovations by RNA to adapt to a changing environment. Some of these adaptations include riboswitches, ribonucleoproteins (RNPs), RNA editing, and RNA interference (RNAi). Although many of these functions may seem

at first glance to be recent evolutionary developments, it may be the case that all of their catalytic activities trace their roots back to a primordial RNA World some four billion years ago, and that RNA's diversity has a continuous thread that pervades life from its very origins. Copyright 2010 John Wiley Sons, Ltd. This article is categorized under:, issn = 1757-7012, doi = 10.1002/wrna.37, url = http:https://doi.org/10.1002/wrna.37, year = 2010, type = Journal Article.

[4] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1):147–57, 1982.

[5] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2):849–57, 1983. Guerrier-Takada, C Gardiner, K Marsh, T Pace, N Altman, S eng Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. 1983/12/01 Cell. 1983 Dec;35(3 Pt 2):849-57.

[6] M. Beringer and M. V. Rodnina. The ribosomal peptidyl transferase. *Mol Cell*, 26(3):311–21, 2007.

[7] Alfonso Mondragn. Structural Studies of RNase P. *Annual Review of Biophysics*, 42(1):537–557, 2013.

[8] John S. Mattick and Igor V. Makunin. Small regulatory RNAs in mammals. *Human Molecular Genetics*, 14:R121–R132, 04 2005.

[9] S. Valadkhan and L. S. Gunawardane. Role of small nuclear rnas in eukaryotic gene expression. *Essays Biochem*, 54:79–90, 2013. Valadkhan, Saba Gunawardane, Lalith S eng Review England 2013/07/09 06:00 Essays Biochem. 2013;54:79-90. doi: 10.1042/bse0540079.

[10] A. D. Garst, A. L. Edwards, and R. T. Batey. Riboswitches: structures and mechanisms. *Cold Spring Harb Perspect Biol*, 3(6), 2011. Garst, Andrew D Edwards, Andrea L Batey, Robert T eng R01 GM073850/GM/NIGMS NIH HHS/ R01 GM073850-07/GM/NIGMS NIH HHS/ R01 GM083953/GM/NIGMS NIH HHS/ R01 GM083953-04/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural Review 2010/10/15 06:00 Cold Spring Harb Perspect Biol. 2011 Jun 1;3(6). pii: cshperspect.a003533. doi: 10.1101/cshperspect.a003533.

[11] Daniel J Klein and Adrian R Ferré-D'Amaré. Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. *Science (New York, N.Y.)*, 313(5794):1752–6, sep 2006.

[12] C. Hammann, A. Luptak, J. Perreault, and M. de la Pena. The ubiquitous hammerhead ribozyme. *RNA*, 18(5):871–85, 2012.

[13] Alan M Lambowitz and Steven Zimmerly. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harbor perspectives in biology*, 3(8):a003616, aug 2011.

[14] T. R. Mercer, M. E. Dinger, and J. S. Mattick. Long non-coding RNAs: insights into functions. *Nat Rev Genet*, 10(3):155–9, 2009.

[15] J. A. Saugstad. Non-Coding RNAs in Stroke and Neuroprotection. *Front Neurol*, 6:50, 2015.

[16] T. Vulliamy, A. Marrone, F. Goldman, A. Dearlove, M. Bessler, P. J. Mason, and I. Dokal. The rna component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature*, 413(6854):432–5, 2001.

[17] R. J. Taft, K. C. Pang, T. R. Mercer, M. Dinger, and J. S. Mattick. Non-coding Rnas: regulators of disease. *J Pathol*, 220(2):126–39, 2010.

[18] T. Gutschner, M. Hammerle, M. Eissmann, J. Hsu, Y. Kim, G. Hung, A. Revenko, G. Arun, M. Stentrup, M. Gross, M. Zornig, A. R. MacLeod, D. L. Spector, and S. Diederichs. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res*, 73(3):1180–9, 2013.

[19] I. Kalvari, J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, R. D. Finn, and A. I. Petrov. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res*, 46(D1):D335–D342, 2018.

[20] J. S. Mattick and I. V. Makunin. Non-coding RNA. *Hum Mol Genet*, 15 Spec No 1:R17–29, 2006.

[21] M. Ridanpaa, H. van Eenennaam, K. Pelin, R. Chadwick, C. Johnson, B. Yuan, W. vanVenrooij, G. Pruijn, R. Salmela, S. Rockas, O. Makitie, I. Kaitila, and A. de la Chapelle. Mutations in the RNA component of

RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia. *Cell*, 104(2):195–203, 2001.

[22] T. Sahoo, D. del Gaudio, J. R. German, M. Shinawi, S. U. Peters, R. E. Person, A. Garnica, S. W. Cheung, and A. L. Beaudet. Prader-willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat Genet*, 40(6):719–21, 2008.

[23] M. Esteller. Non-coding rnas in human disease. *Nat Rev Genet*, 12(12):861–74, 2011.

[24] Y. Qi, H. S. Ooi, J. Wu, J. Chen, X. Zhang, S. Tan, Q. Yu, Y. Y. Li, Y. Kang, H. Li, Z. Xiong, T. Zhu, B. Liu, Z. Shao, and X. Zhao. MALAT1 long ncRNA promotes gastric cancer metastasis by suppressing PCDH10. *Oncotarget*, 7(11):12693–703, 2016.

[25] J. A. Howe, H. Wang, T. O. Fischmann, C. J. Balibar, L. Xiao, A. M. Galgoci, J. C. Malinverni, T. Mayhood, A. Villafania, A. Nahvi, N. Murgolo, C. M. Barbieri, P. A. Mann, D. Carr, E. Xia, P. Zuck, D. Riley, R. E. Painter, S. S. Walker, B. Sherborne, R. de Jesus, W. Pan, M. A. Plotkin, J. Wu, D. Rindgen, J. Cummings, C. G. Garlisi, R. Zhang, P. R. Sheth, C. J. Gill, H. Tang, and T. Roemer. Selective small-molecule inhibition of an RNA structural element. *Nature*, 526(7575):672–7, 2015.

[26] D. N. Wilson. Ribosome-targeting antibiotics and mechanisms of bacterial resistance. *Nat Rev Microbiol*, 12(1):35–48, 2014.

[27] U. Schmitz, H. Naderi-Meshkin, S. K. Gupta, O. Wolkenhauer, and J. Vera. The RNA world in the 21st century-a systems approach to finding non-coding keys to clinical questions. *Brief Bioinform*, 2015.

[28] K. M. Weeks. Advances in RNA structure analysis by chemical probing. *Curr Opin Struct Biol*, 20(3):295–304, 2010.

[29] N. R. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, 453:3–31, 2008.

[30] M. Parisien and F. Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–5, 2008.

[31] J. Reeder, M. Hochsmann, M. Rehmsmeier, B. Voss, and R. Giegerich. Beyond Mfold: recent advances in RNA bioinformatics. *J Biotechnol*, 124(1):41–55, 2006.

[32] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–15, 2003.

[33] Z. Miao and E. Westhof. RNA Structure: Advances and Assessment of 3D Structure Prediction. *Annual Review of Biophysics*, 46(1):483–503, 2017.

[34] J. A. Nelson and O. C. Uhlenbeck. Hammerhead redux: does the new structure fit the old biochemical data? *RNA*, 14(4):605–15, 2008.

[35] A. Wilkins, S. Erdin, R. Lua, and O. Lichtarge. Evolutionary trace for prediction and redesign of protein functional sites. *Methods Mol Biol*, 819:29–42, 2012.

[36] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal. Consurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*, 38(Web Server issue):W529–33, 2010.

[37] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, 6(12):e1001025, 2010.

[38] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–50, 2005.

[39] M. Levitt. Detailed molecular model for transfer ribonucleic acid. *Nature*, 224(5221):759–63, 1969.

[40] G. E. Fox and C. R. Woese. 5S RNA secondary structure. *Nature*, 256(5517):505–7, 1975.

[41] F. Michel and E. Westhof. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol*, 216(3):585–610, 1990.

[42] Caleb Weinreb, Adam J. Riesselman, John B. Ingraham, Torsten Gross, Chris Sander, and Debora S. Marks. 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell*, 165(4):963–975, 2016.

[43] O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2):342–58, 1996.

[44] I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol*, 336(5):1265–82, 2004.

[45] P. Katsonis and O. Lichtarge. A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. *Genome Res*, 24(12):2050–8, 2014.

[46] O. Lichtarge, H. Yao, D. M. Kristensen, S. Madabushi, and I. Mihalek. Accurate and scalable identification of functional sites by evolutionary tracing. *J Struct Funct Genomics*, 4(2-3):159–66, 2003.

[47] L. Rajagopalan, N. Patel, S. Madabushi, J. A. Goddard, V. Anjan, F. Lin, C. Shope, B. Farrell, O. Lichtarge, A. L. Davidson, W. E. Brownell, and F. A. Pereira. Essential helix interactions in the anion transporter domain of prestin revealed by evolutionary trace analysis. *J Neurosci*, 26(49):12727–34, 2006.

[48] I. Res, I. Mihalek, and O. Lichtarge. An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, 21(10):2496–501, 2005.

[49] G. J. Rodriguez, R. Yao, O. Lichtarge, and T. G. Wensel. Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proc Natl Acad Sci U S A*, 107(17):7787–92, 2010.

[50] M. Raviscioni, Q. He, E. M. Salicru, C. L. Smith, and O. Lichtarge. Evolutionary identification of a subtype specific functional site in the ligand binding domain of steroid receptors. *Proteins*, 64(4):1046–57, 2006.

[51] P. Gu, D. H. Morgan, M. Sattar, X. Xu, R. Wagner, M. Raviscioni, O. Lichtarge, and A. J. Cooney. Evolutionary trace-based peptides identify a novel asymmetric interaction that mediates oligomerization in nuclear receptors. *J Biol Chem*, 280(36):31818–29, 2005.

[52] S. Erdin, E. Venner, A. M. Lisewski, and O. Lichtarge. Function prediction from networks of local evolutionary similarity in protein structure. *BMC Bioinformatics*, 14 Suppl 3:S6, 2013.

[53] D. M. Kristensen, R. M. Ward, A. M. Lisewski, S. Erdin, B. Y. Chen, V. Y. Fofanov, M. Kimmel, L. E. Kavraki, and O. Lichtarge. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics*, 9:17, 2008.

[54] S. Madabushi, H. Yao, M. Marsh, D. M. Kristensen, A. Philippi, M. E. Sowa, and O. Lichtarge. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J Mol Biol*, 316(1):139–54, 2002.

[55] I. Mihalek, I. Res, and O. Lichtarge. Evolutionary and structural feedback on selection of sequences for comparative analysis of proteins. *Proteins*, 63(1):87–99, 2006.

[56] A. D. Wilkins, E. Venner, D. C. Marciano, S. Erdin, B. Atri, R. C. Lua, and O. Lichtarge. Accounting for epistatic interactions improves the functional analysis of protein structures. *Bioinformatics*, 29(21):2714–21, 2013.

[57] I. Mihalek, I. Res, H. Yao, and O. Lichtarge. Combining inference from evolution and geometric probability in protein structure evaluation. *J Mol Biol*, 331(1):263–79, 2003.

[58] M. J. Fedor. Comparative enzymology and structural biology of rna self-cleavage. *Annu Rev Biophys*, 38:271–99, 2009.

[59] Y. I. Chi, M. Martick, M. Lares, R. Kim, W. G. Scott, and S. H. Kim. Capturing hammerhead ribozyme structures in action by modulating general base catalysis. *PLoS Biol*, 6(9):e234, 2008.

[60] A. Khvorova, A. Lescoute, E. Westhof, and S. D. Jayasena. Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. *Nat Struct Biol*, 10(9):708–12, 2003.

[61] A. Mir, J. Chen, K. Robinson, E. Lendy, J. Goodman, D. Neau, and B. L. Golden. Two divalent metal ions and conformational changes play roles in the hammerhead ribozyme cleavage reaction. *Biochemistry*, 54(41):6369–81, 2015.

[62] C. Hammann, D. G. Norman, and D. M. Lilley. Dissection of the ion-induced folding of the hammerhead ribozyme using 19f nmr. *Proc Natl Acad Sci U S A*, 98(10):5503–8, 2001.

[63] N. Demeshkina, L. Jenner, E. Westhof, M. Yusupov, and G. Yusupova. A new understanding of the decoding principle on the ribosome. *Nature*, 484(7393):256–9, 2012.

[64] A. S. Petrov, C. R. Bernier, C. Hsiao, A. M. Norris, N. A. Kovacs, C. C. Waterbury, V. G. Stepanov, S. C. Harvey, G. E. Fox, R. M. Wartell, N. V. Hud, and L. D. Williams. Evolution of the ribosome at atomic resolution. *Proc Natl Acad Sci U S A*, 111(28):10251–6, 2014.

[65] R. M. Voorhees, A. Weixlbaumer, D. Loakes, A. C. Kelley, and V. Ramakrishnan. Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70s ribosome. *Nat Struct Mol Biol*, 16(5):528–33, 2009.

[66] M. Nomura. Assembly of bacterial ribosomes. *Science*, 179(4076):864–73, 1973.

[67] Q. Liu and K. Fredrick. Intersubunit bridges of the bacterial ribosome. *J Mol Biol*, 428(10 Pt B):2146–64, 2016.

[68] J. A. Kowalak, E. Bruenger, and J. A. McCloskey. Posttranscriptional modification of the central loop of domain v in escherichia coli 23 s ribosomal rna. *J Biol Chem*, 270(30):17758–64, 1995.

[69] W. Krzyzosiak, R. Denman, K. Nurse, W. Hellmann, M. Boublik, C. W. Gehrke, P. F. Agris, and J. Ofengand. In vitro synthesis of 16s ribosomal rna containing single base changes and assembly into a functional 30s ribosome. *Biochemistry*, 26(8):2353–64, 1987.

[70] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Muller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell. The comparative rna web (crw) site: an online database of comparative sequence and structure information for ribosomal, intron, and other rnas. *BMC Bioinformatics*, 3:2, 2002.

[71] B. P. Klaholz, A. G. Myasnikov, and M. Van Heel. Visualization of release factor 3 on the ribosome during termination of protein synthesis. *Nature*, 427(6977):862–5, 2004.

[72] P. Julian, P. Milon, X. Agirrezabala, G. Lasso, D. Gil, M. V. Rodnina, and M. Valle. The cryo-em structure of a complete 30s translation initiation complex from escherichia coli. *PLoS Biol*, 9(7):e1001095, 2011.

[73] I. Mihalek, I. Res, and O. Lichtarge. A structure and evolution-guided monte carlo sequence selection strategy for multiple alignment-based analysis of proteins. *Bioinformatics*, 22(2):149–56, 2006.

[74] S. Kim, N. K. Yu, and B. K. Kaang. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. Mol. Med.*, 47:e166, Jun 2015.

[75] A. Bruckner, C. Polge, N. Lentze, D. Auerbach, and U. Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci*, 10(6):2763–2788, Jun 2009.

[76] F. X. Sutandy, J. Qian, C. S. Chen, and H. Zhu. Overview of protein microarrays. *Curr Protoc Protein Sci*, Chapter 27:Unit 27.1, Apr 2013.

[77] H. Iqbal, D. R. Akins, and M. R. Kenedy. Co-immunoprecipitation for Identifying Protein-Protein Interactions in Borrelia burgdorferi. *Methods Mol. Biol.*, 1690:47–55, 2018.

[78] A. Ilari and C. Savino. Protein structure determination by x-ray crystallography. *Methods Mol. Biol.*, 452:63–87, 2008.

[79] M. Winey, J. B. Meehl, E. T. O'Toole, and T. H. Giddings. Conventional transmission electron microscopy. *Mol. Biol. Cell*, 25(3):319–323, Feb 2014.

[80] J. H. Bothwell and J. L. Griffin. An introduction to biological nuclear magnetic resonance spectroscopy. *Biol Rev Camb Philos Soc*, 86(2):493–510, May 2011.

[81] UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, 45(D1):D158–D169, 01 2017.

[82] Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, D. D'Andrea, R. Lepore, C. S. Funk, I. Kahanda, K. M. Verspoor, A. Ben-Hur, d. a. C. E. Koo, D. Penfold-Brown, D. Shasha, N. Youngs, R. Bonneau, A. Lin, S. M. Sahraeian, P. L. Martelli, G. Profiti, R. Casadio, R. Cao, Z. Zhong, J. Cheng, A. Altenhoff, N. Skunca, C. Dessimoz, T. Dogan, K. Hakala, S. Kaewphan, F. Mehryary, T. Salakoski, F. Ginter, H. Fang, B. Smithers, M. Oates, J. Gough, P. Toronen, P. Koskinen, L. Holm, C. T. Chen, W. L. Hsu, K. Bryson, D. Cozzetto, F. Minneci, D. T. Jones, S. Chapman, D. Bkc, I. K. Khan, D. Kihara, D. Ofer, N. Rappoport, A. Stern, E. Cibrian-Uhalte, P. Denny, R. E. Foulger, R. Hieta, D. Legge, R. C. Lovering, M. Magrane, A. N. Melidoni, P. Mutowo-Meullenet, K. Pichler, A. Shypitsyna, B. Li, P. Zakeri, S. ElShal, L. C. Tranchevent, S. Das, N. L. Dawson, D. Lee, J. G. Lees, I. Sillitoe, P. Bhat, T. Nepusz, A. E. Romero, R. Sasidharan, H. Yang, A. Paccanaro, J. Gillis, A. E. Sedeno-Cortes, P. Pavlidis, S. Feng, J. M. Cejuela, T. Goldberg, T. Hamp, L. Richter, A. Salamov, T. Gabaldon, M. Marcet-Houben, F. Supek, Q. Gong, W. Ning, Y. Zhou, W. Tian, M. Falda, P. Fontana, E. Lavezzo, S. Toppo, C. Ferrari, M. Giollo, D. Piovesan, S. C. Tosatto, A. Del Pozo, J. M. Fernandez, P. Maietta, A. Valencia, M. L. Tress, A. Benso, S. Di Carlo, G. Politano, A. Savino, H. U. Rehman, M. Re, M. Mesiti, G. Valentini, J. W. Bargsten, A. D. van Dijk, B. Gemovic, S. Glisic, V. Perovic, V. Veljkovic, N. Veljkovic, D. C. Almeida-E-Silva, R. Z.

Vencio, M. Sharan, J. Vogel, L. Kansakar, S. Zhang, S. Vucetic, Z. Wang, M. J. Sternberg, M. N. Wass, R. P. Huntley, M. J. Martin, C. O'Donovan, P. N. Robinson, Y. Moreau, A. Tramontano, P. C. Babbitt, S. E. Brenner, M. Linial, C. A. Orengo, B. Rost, C. S. Greene, S. D. Mooney, I. Friedberg, and P. Radivojac. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, 17(1):184, 09 2016.

[83] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990.

[84] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, Sep 1997.

[85] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta. Pfam: the protein families database. *Nucleic Acids Res.*, 42(Database issue):D222–230, Jan 2014.

[86] L. Holm, S. Kaariainen, C. Wilton, and D. Plewczynski. Using Dali for structural comparison of proteins. *Curr Protoc Bioinformatics*, Chapter 5:Unit 5.5, Jul 2006.

[87] C. A. Orengo and W. R. Taylor. SSAP: sequential structure alignment program for protein structure comparison. *Meth. Enzymol.*, 266:617–635, 1996.

[88] C. Zhang, P. L. Freddolino, and Y. Zhang. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res.*, 45(W1):W291–W299, Jul 2017.

[89] N. L. Dawson, T. E. Lewis, S. Das, J. G. Lees, D. Lee, P. Ashford, C. A. Orengo, and I. Sillitoe. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.*, 45(D1):D289–D295, 01 2017.

[90] C. J. Sigrist, L. Cerutti, E. de Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, and N. Hulo. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, 38(Database issue):D161–166, Jan 2010.

[91] S. Erdin, E. Venner, A. M. Lisewski, and O. Lichtarge. Function prediction from networks of local evolutionary similarity in protein structure. *BMC Bioinformatics*, 14 Suppl 3:S6, 2013.

[92] J. Li, S. K. Halgamuge, C. I. Kells, and S. L. Tang. Gene function prediction based on genomic context clustering and discriminative learning: an application to bacteriophages. *BMC Bioinformatics*, 8 Suppl 4:S6, May 2007.

[93] M. Kulmanov, M. A. Khan, R. Hoehndorf, and J. Wren. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, 02 2018.

[94] A. M. Lisewski, J. P. Quiros, C. L. Ng, A. K. Adikesavan, K. Miura, N. Putluri, R. T. Eastman, D. Scanfeld, S. J. Regenbogen, L. Altenhofen, M. Llinas, A. Sreekumar, C. Long, D. A. Fidock, and O. Lichtarge. Supergenomic network compression and the discovery of EXP1 as a glutathione transferase inhibited by artesunate. *Cell*, 158(4):916–928, Aug 2014.

[95] Scott Spangler, Angela D. Wilkins, Benjamin J. Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R. Pickering, Austin Comer, Jeffrey N. Myers, Ioana Stanoi, Linda Kato, Ana Lelescu, Jacques J. Labrie, Neha Parikh, Andreas Martin Lisewski, Lawrence Donehower, Ying Chen, and Olivier Lichtarge. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1877–1886, New York, NY, USA, 2014. ACM.

[96] B. K. Choi, T. Dayaram, N. Parikh, A. D. Wilkins, M. Nagarajan, I. B. Novikov, B. J. Bachman, S. Y. Jung, P. J. Haas, J. L. Labrie, C. R. Pickering, A. K. Adikesavan, S. Regenbogen, L. Kato, A. Lelescu, C. M. Buchovecky, H. Zhang, S. H. Bao, S. Boyer, G. Weber, K. L. Scott, Y. Chen, S. Spangler, L. A. Donehower, and O. Lichtarge. Literature-based automated discovery

of tumor suppressor p53 phosphorylation and inhibition by NEK2. *Proc. Natl. Acad. Sci. U.S.A.*, 115(42):10666–10671, 10 2018.

[97] No authors listed. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, 47(D1):D330–D338, Jan 2019.

[98] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001.

[99] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, 5(7):e1000443, Jul 2009.

[100] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schlkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pages 321–328. MIT Press, 2004.

[101] H. Wang, H. Huang, C. Ding, and F. Nie. Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. *J. Comput. Biol.*, 20(4):344–358, Apr 2013.

[102] S. Ray and S. Bandyopadhyay. A NMF based approach for integrating multiple data sources to predict HIV-1-human PPIs. *BMC Bioinformatics*, 17:121, Mar 2016.

[103] S. Regenbogen, A. D. Wilkins, and O. Lichtarge. COMPUTING THERAPY FOR PRECISION MEDICINE: COLLABORATIVE FILTERING INTEGRATES AND PREDICTS MULTI-ENTITY INTERACTIONS. *Pac Symp Biocomput*, 21:21–32, 2016.

[104] S. Jain and G. D. Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, 11(1):562, Nov 2010.

[105] S. B. Zhang and Q. R. Tang. Protein-protein interaction inference based on semantic similarity of Gene Ontology terms. *J. Theor. Biol.*, 401:30–37, 07 2016.

[106] Catia Pesquita, Daniel Faria, Hugo Bastos, António EN Ferreira, André O. Falcão, and Francisco M. Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(5):S4, Apr 2008.

[107] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *CoRR*, abs/1105.5444, 2011.

[108] P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, and E. Skrzypek. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, 43(Database issue):D512–520, Jan 2015.

[109] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, 37(Database issue):D793–796, Jan 2009.

[110] Eva Freyhult, Vincent Moulton, and Paul Gardner. Predicting RNA Structure Using Mutual Information. *Applied Bioinformatics*, 4(1):53–59, Mar 2005.

[111] Song Cao and Shi-Jie Chen. Physics-Based De Novo Prediction of RNA 3D Structures. *The Journal of Physical Chemistry B*, 115(14):4216–4226, 2011.