Detection of Non-coding RNA with Comparative Genomics and the Sequential Closure of Smooth Graphs in Cartesian Currents

by

Alex Coventry

BSc. (Hons) Australian National University, 1994

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
February 2003

Author . ...
Department of Mathematics
January 21, 2003

Certified by.. ....
Bonnie Berger
Professor of Applied Mathematics
Thesis Supervisor

Accepted by.. ....
Rodolfo Ruben Rosales
Chairman, Applied Mathematics Committee

Accepted by.... ....
Pavel Etingof
Chairman, Department Committee on Graduate Students

Detection of Non-coding RNA with Comparative Genomics and the Sequential Closure of Smooth Graphs in Cartesian Currents

by

Alex Coventry

Submitted to the Department of Mathematics
on January 21, 2003, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In the field of genomics, this thesis presents algorithms for identifying non-coding RNA (ncRNA) genes. It describes a rapid and highly reliable comparative statistical method for identification of functionally significant base pairs in ncRNA genes in multiple sequence alignments of cross-species homologs, a divide-and-conquer approach to optimal assembly of exon predictions with $O(n \log n)$ time-complexity, (the standard algorithm for exon assembly has $O(n^2)$ time-complexity for ncRNA exon predictions,) and highly accurate statistical tests for exon boundaries based on recognition of non-contiguous patterns in known examples. It also describes a method for scanning cDNA for ncRNA genes.

In the field of geometric measure theory, it proves that the set of cartesian currents given by integration over the graphs of smooth functions is dense in the set of all cartesian currents.

Thesis Supervisor: Bonnie Berger
Title: Professor of Applied Mathematics

Acknowledgments

I'm very grateful to my advisor, Bonnie Berger, for her patience, support, advice and encouragement of my research. She has given me many opportunities to work with biologists, and free rein in the direction of my research. My greatest progress came when my funding ran out, so I am especially grateful for that. I've also had helpful discussions about my research with Danny Kleitman, during which he's given me lots of good ideas, as well as a feel for interesting research topics. Our discussions may have played a role in the development of the methods described in section 2.2.3, which are reminiscent of techniques mentioned in passing in [42]. Discussions with my undergraduate advisor, John Hutchinson, were tremendously helpful in the research described in chapter 5. It was largely John's encouragement of my undergraduate research that led me to specialize in Mathematics, and he helped me refine my ideas about cartesian currents even after I came to MIT.

I'm also grateful to Monami Chakrabarti, who's been unstintingly supportive, even when I've had to hide away from everyone to complete my research. My parents and siblings have also been very encouraging when things have looked bleak, and have also helped me in numerous other ways.

My friends in SIPB have made the SIPB office a pleasant and occasionally productive place to work, and I've learned a lot from them about software development and general computational issues that has directly assisted me in the research I describe here.

I'd also like to thank Linda Okun, who was very sympathetic when I was floundering, and patiently helped me with Institute administrative issues for years.

My first few years at MIT were really tough, and a lot of people showed me a great deal of kindness then. Thanks to all of you.

Cambridge, Massachusetts
Alex Coventry
January 21, 2003

Contents

Chapter 1

Introduction

1.1 Gene recognition

In principle, it should be possible to infer an organism's biology almost entirely from its genome, as its genes specify the vast majority of its chemical constituents. At this stage, however, not only is there no way to predict how those chemicals will behave, but many organisms' genomes (including humans') represent genes themselves so cryptically that currently they are only imperfectly discernible without painstaking *in vitro* simulation. The mechanisms of gene expression are fairly well understood at a gross level [34], but they involve affinities between proteins and DNA or RNA subsequences that are very hard to identify reliably from gene sequences. As a result, even though the human genome has been sequenced [11], it almost certainly contains genes that have not yet been identified. There is simply too much of it to exhaustively search for genes using existing lab techniques.

Significant progress in identifying genes computationally has been made using statistical analyses of known genes [8, 9], but by themselves, these tests are not accurate enough for a genome-scale search.

To improve on the accuracy of computational gene prediction, searches for genes have turned to other genomic structure. In the past few years, *comparative genomics* has become a popular approach [4, 54]. Its key idea is that usually the rate of mutation in biologically significant regions is strongly retarded by evolutionary selection pressure. Thus searching genomes of multiple organisms for strongly similar regions and restricting the search for genes to those regions dramatically winnows the set of candidate genes, sometimes to the extent that existing computational methods suffice to identify genes reasonably accurately. Comparative genomics can be further refined by scoring similar regions according to the extent to which the variations between them match the expected mutation patterns for genes [41].

Nearly all gene finding efforts have concentrated on genes that generate proteins [8, 9, 4, 50, 19]. However, a lot of RNA transcribed from DNA is never translated to protein, and some of it, such as the XIST gene, which in males suppresses expression of genes on the X chromosome [15], is biologically essential. Identifying such genes will be crucial to a complete understanding of cellular biology.

There appears to have been relatively little work on automatic identification of non-coding RNA (ncRNA [47]) genes. Some approaches have focused on searching for a recognizable secondary structure associated with RNA transcripts serving a specific biological function. An example is Regalia et al.'s [44] search for signal recognition particles and Rhoades et al.'s. search for microRNA's [45]. I am aware of two programs that search for generic traits of ncRNA genes. One is RNAGENiE [10], which combines searches for motifs known to occur in RNA genes, free-energy calculations, and neural-nets and support-vectors trained on known examples. The other is QRNA [47], which scans pairwise alignments of homologous DNA sequences from related genomes, and flags homologs that exhibit mutation patterns concomitant with selection-driven preservation of RNA secondary structure.

In this thesis, I present new algorithms for detecting ncRNA genes. I describe the first ncRNA detection scheme using multiple sequence alignments (MSA's) of more than two homologs. By exploiting the extra information in a large MSA, this algorithm outstrips QRNA in terms of both speed and accuracy. MSA's have been used for determination of ncRNA secondary structure for decades [27, 18, 30, 43] , but not for searching for ncRNA genes. As I demonstrate with a scan of the Buchnera genome against eighteen other bacterial genomes, this is becoming a feasible approach.

The key idea of the algorithm is to estimate the statistical significance of the number of reverse-complementary portions of an MSA's constituent sequences. In the first phases of the algorithm's development, this estimate was based on searching for complementary base pairs at each pair of positions in an MSA. This is similar to the manual approach commonly used to predict RNA secondary structure through comparative analysis. Automated thermodynamics-based secondary structure prediction has been augmented with similar comparative analysis of associated MSA's by Hofacker et al. [27], but they use a fixed column-based score to detect conservation of base pairs, and are dependant on correct alignment of orthologous base pairs in the MSA's they use.

At the time of writing, the latest version of the algorithm is based on identifying potential helices using a Smith-Waterman-style alignment of sequences in the MSA to their reverse complements. By allowing for gaps and mismatches, this method can detect reverse-complementary regions which are very significant, and this permits even greater tolerance of misalignment of orthologous base pairs within the MSA.

The runtime of all algorithms is $O(n^2)$ in the length of the MSA, so they are feasible for larger-scale searches than QRNA, the run time of which grows cubically in the length of the alignment it is passed.

A scan based on MSA's also has the advantage of often working even if some of its constituent alignments are too weak for pairwise alignment to capture significant features, or too strong to exhibit significant covariation in complementary base pairs.

In this thesis, I also describe a comparative-genomics algorithm similar to ROSETTA [4] incorporating a new exon assembly algorithm of time-complexity $O(N \log N)$, and new splice-site recognition algorithms based on richer statistical models than existing methods use, taking advantage of the huge number of example splice sites now available. I also describe a PROCRUSTES-like [19] approach to scanning cDNA's for ncRNA genes. This involves BLAST'ing [2] cDNA's from one organism against the genome of another, and concatenating appropriately ordered matches into a single alignment that can be tested using QRNA.

1.2 Geometric Measure Theory

In chapter 5, I describe a proof of a long standing conjecture regarding Cartesian Currents. My investigation of this conjecture began while I was an undergraduate, and it was then that I devised the constructions used in lemma 5.3.1, which show $\mathsf{Cart}(\Omega, \mathbb{R}^N)$ to be closed under arbitrary modifications to the "vertical" portion of its elements. However, I realized the key dimension-reduction technique described in section 5.4 while at MIT.

Let $\mathcal{F} : C^\infty(\Omega, \mathbb{R}^N) \to \mathbb{R}$ be a polyconvex functional such that for some fixed $c > 0$ and for $u \in C^\infty(\Omega, \mathbb{R}^N)$, $\mathcal{F}(u) \geq c\mathcal{H}^n(G_u)$, where G_u is the graph of u, $\Omega \subset \mathbb{R}^N$ an open set with smooth boundary and $\mathcal{H}^n$ is the n-dimensional Hausdorff measure. Giaquinta, Modica and Souček [21] introduced the space $\mathsf{cart}(\Omega, \mathbb{R}^N)$ (see Definition 5.1.2) as a natural space in which to seek minimizers for $\mathcal{F}$. Roughly, an element of $\mathsf{cart}(\Omega, \mathbb{R}^N)$ is an n-current given by integration over the union of the graph of a function $u : \Omega \to \mathbb{R}^N$ (called the underlying function of the current) and a "vertical part". An element of this space is called a *cartesian current*. A sequence of cartesian currents is said to converge C-weakly if it converges weakly, its elements have uniformly bounded mass, and their underlying functions have uniformly bounded L^1-norm.

Let $\mathsf{Cart}(\Omega, \mathbb{R}^N) \subset \mathsf{cart}(\Omega, \mathbb{R}^N)$ denote the smallest subset which is closed under C-weak convergence and contains all currents which are given by integration over the graphs of smooth functions from Ω to $\mathbb{R}^N$. In [20], Giaquinta, Modica and Souček suggested that the conjecture $\mathsf{Cart}(\Omega, \mathbb{R}^N) = \mathsf{cart}(\Omega, \mathbb{R}^N)$ seems reasonable. The main result of this chapter is that this is indeed the case when Ω is a smooth bounded domain.

Giaquinta, Modica and Souček [20] discussed the Dirichlet Energy and the phenomenon of "bubbling of spheres" in terms of cartesian currents. They characterized the weak C-limits of sequences of cartesian currents given by integration over graphs of functions with uniformly bounded Dirichlet Energy from Ω to $\mathrm{S}^2 \subset \mathbb{R}^3$.

They also considered the following more general situation: for $p > 1$ and open $\Omega \subset$

$\mathbb{R}^n$, let $\mathcal{A}^p(\Omega, \mathbb{R}^N)$ be the subset of $L^p(\Omega, \mathbb{R}^N)$ comprised of the weakly differentiable functions whose weak derivatives have minors in $L^p(\Omega, \mathbb{R}^N)$. A sequence (u_k) is said to converge weakly in $\mathcal{A}^p$ if the functions (u_k) and all the minors of their derivatives converge L^p-weakly. Let $\mathsf{Cart}^p(\Omega, \mathbb{R}^N)$ be the smallest set containing $C^1(\Omega, \mathbb{R}^N) \cap \mathcal{A}^p(\Omega, \mathbb{R}^N)$ which is closed under this notion of convergence, and let $\mathsf{cart}^p(\Omega, \mathbb{R}^N)$ be the set of all functions $u \in \mathcal{A}^p(\Omega, \mathbb{R}^N)$ such that the current given by integration over the graph of u is boundaryless. They raised the question of whether $\mathsf{cart}^p(\Omega, \mathbb{R}^N) = \mathsf{Cart}^p(\Omega, \mathbb{R}^N)$, and whether every element of $\mathsf{cart}^p(\Omega, \mathbb{R}^N)$ is the limit of graphs of smooth functions. Malý showed that this is not the case in [35].

In this thesis, the proof that for smooth bounded Ω, $\mathsf{Cart}(\Omega, \mathbb{R}^N) = \mathsf{cart}(\Omega, \mathbb{R}^N)$ uses induction on Ω's dimension. Firstly, Lemmas 5.5.1 and 5.5.2 show that

$$\mathcal{F} = \{T \in \mathsf{cart}(\Omega, \mathbb{R}^N) \mid \text{spt } (T - [\![G_0]\!]) \subset\subset \Omega \times \mathbb{R}^N\}$$

is dense in $\mathsf{cart}(\Omega, \mathbb{R}^N)$. Then Corollary 5.4.4 shows that for T in this set, there is a sequence of currents $S_k \overset{\mathsf{C}}{\rightharpoonup} T$ which are constructed in a very explicit fashion from cartesian currents of dimension one less. Specifically, Lemma 5.4.1 shows that for a sequence $\rho_k \to 0$, there is a grid of tessellating open n-cubes $\{A_{k,l} \subset \Omega\}$ of side-length ρ_k and centers $q_{k,l}$ such that slicing T by $\bigcup \partial A_{k,l} \times \mathbb{R}^N$ gives a current whose restriction to any n-dimensional linear subspace of the grid is a cartesian current. If $p_{A_{k,l}} : (x, y) \in \Omega \times \mathbb{R}^N \mapsto (q_{k,l}, y) \in \{q_{k,l}\} \times \mathbb{R}^N$ and $h_{A_{k,l}}$ is the affine homotopy of this map with the identity, then

$$S_k = \Sigma_l h_{A_{k,l}\#}([\![0,1]\!] \times \partial(T\lfloor A_{k,l})) + p_{A_{k,l}\#}(T\lfloor A_{k,l}) + [\![(\Omega \setminus \bigcup_l A_{k,l}) \times \{0\}]\!]$$

is a weak approximation to T. Finally, in Theorem 5.4.4 the explicit expression of S in terms of $(n-1)$-dimensional currents is combined with the induction hypothesis and corollary 5.2.3 to show that $S_k \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$. Hence $T \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$, completing the proof.

Whether the conjecture holds for arbitrary domains is still unclear. A crucial step in the proof is showing that $\mathcal{F}$ is dense in $\mathsf{cart}(\Omega, \mathbb{R}^N)$, and there is no straightforward way to duplicate the proof of this for arbitrary open sets.

Chapter 2

Prediction of non-coding RNA (ncRNA) genes using comparative sequence analysis

This chapter describes new comparative techniques for identification of orthologous genes whose function depends on secondary structure in the RNA transcripts they generate.

Eventually understanding exactly how cells find and express genes is important, but merely identifying genes is of enormous value in itself. Since we can draw inferences from the genomes of multiple organisms, using information that is unavailable to cellular expression machinery, many genes can be found without a perfect understanding of how they are expressed.

Any set of organisms has a common ancestor, and usually most of their genes are descended from a gene in that ancestor. Genes in different species that are believed to have descended from a common ancestral gene are called *orthologs.* Genes are delicate, and random mutations of them often weaken the resulting organism, reducing the likelihood that they will have descendants to pass the mutations on to. Thus the rate of mutation from generation to generation tends to be much lower in genes than in portions of the genome that are never expressed, and one way to narrow the search for genes is to look for them in regions of genomes that are strongly similar, then search for orthologs in these regions.

Unfortunately, evolutionary pressure on the phenotype is not the only mechanism for preservation of DNA sequences across generations—for instance, the fidelity with which splice sites and transcription start sites are preserved varies a lot, and is often very high, in some cases almost perfect, even far from the boundary that they signal. Also, there is a wide variation in the fidelity with which orthologous genes match, presumably corresponding to variation in the delicacy and importance of the genes' products. Thus, while comparative methods can winnow the field of candidate genes, they do not yet reduce accurate gene prediction to a matter of searching for similar strings. For the time being, they have to be used in conjunction with tests for local features such as those described in chapter 4, as imperfect as those are.

2.1 Non-coding RNA genes

The role of RNA in the early "central dogma" of cell biology was limited to mRNA transcripts, tRNA's and rRNA's. However there are other RNA sequences known to be biologically functional, such as RNaseP [7], SRP [25], snoRNA and XIST [15]. Eddy's review, "Non-coding RNA genes and the modern RNA world" [15], gives a good overview of the current state of research into ncRNA.

2.1.1 RNA secondary structure

The functions of biologically significant RNA molecules often depend on their *secondary structure*. As with DNA, nucleotides in RNA molecules can base-pair with each other, and the nucleotides of single strands of RNA can base-pair to form complex structures. The set of such pairs are referred to as a molecule's secondary structure, to distinguish it from its *tertiary structure*—its three-dimensional geometry. The nucleotides in RNA are usually represented with the symbols `A`, `C`, `G` and `U`. These are transcribed from the DNA nucleotides `A`, `C`, `G` and `T` respectively. At times, `U` and `T` may be used interchangeably in this thesis. The stable base pairs in RNA are `A-U`, `G-U` and `G-C`. Note that this is one more possible pairing than in DNA.

abcdef.....ghijk....lmn....nml....kjihg.fedcba..
CCCCCa-aacc-CCGcuaggUCCggaaGGAagcaaCGGu-aGGGGGac
gUCGcc-aacc-CGGUcaggUCCggaaGGAagcaGCCGu-aaCGAauu
CUCGcc-aacc-UGGUcaggGCCgagaGGCagcaGCCAc-aaCGAGau
UCUUGCuuag-UUGGUcaggUCUgaaaAGAagcaGCCAGgGUAAGAuu
cCCAUg-aacc-UGGUcaggUCCggaaGGAagcaGCCAuaaGUGGauc
cuUCUg-aacc-GGGUcaggAUCggaaGGUagcaGCCCuaaGGAuagg
AUUGCUgaauc-CCGUcaggACUggaaGGUagcaGCGGuaAGCGAUuu
AUUGUg-aacc-CCGUcaggCCCggaaGGGagcaGCGGua-GCAGUug
CCCGUc-aacc-UGGUcaggUCCggaaGGAagcaGCCAca-GCGGGaa

Table 2.1: MSA of SRP RNA from [25]. Each row contains a portion of an SRP gene's nucleotide sequence, with some gaps included for alignment purposes. In the top row, pairs of letters indicate positions at which some of the SRP orthologs are believed to base pair. Nucleotides in a sequence which are believed to base pair are in captitals. This alignment of SRP's is a portion of the alignment currently available from `http://psyche.uthct.edu/dbs/SRPDB/rna/alignment/text/srprna_ali.80coltext`.

For instance, snoRNA's have the behavior and associated secondary structure shown in figure 2-1.

2.1.2 Existing methods for RNA secondary structure prediction

Secondary structures of orthologous ncRNA's have been determined by hand using comparative techniques for many years (see, e.g., [18]) but have only recently been used to detect them [47]. The key idea in comparative analysis of ncRNA's is that functionally significant secondary structure encourages distinctive patterns in variations among orthologs. Table 2.1 contains an example of such patterns. It shows a portion of a multiple sequence alignment (MSA) of bacterial SRP [25] genes.

The pairs of symbols "a", "b", "c"... in the first row indicate columns that are believed to be paired to each other in the consensus SRP secondary structure. Nucleotides in a sequence that are believed to be base-paired are indicated in capital letters. Thus in the "c" columns, the consensus nucleotides are C and G respectively, and it is believed that these are base paired. In rows where this consensus is violated, their complementarity is maintained. Thus, in a row with an A in one "c" column, the other "c" column has a U. The secondary structure of orthologs has been determined

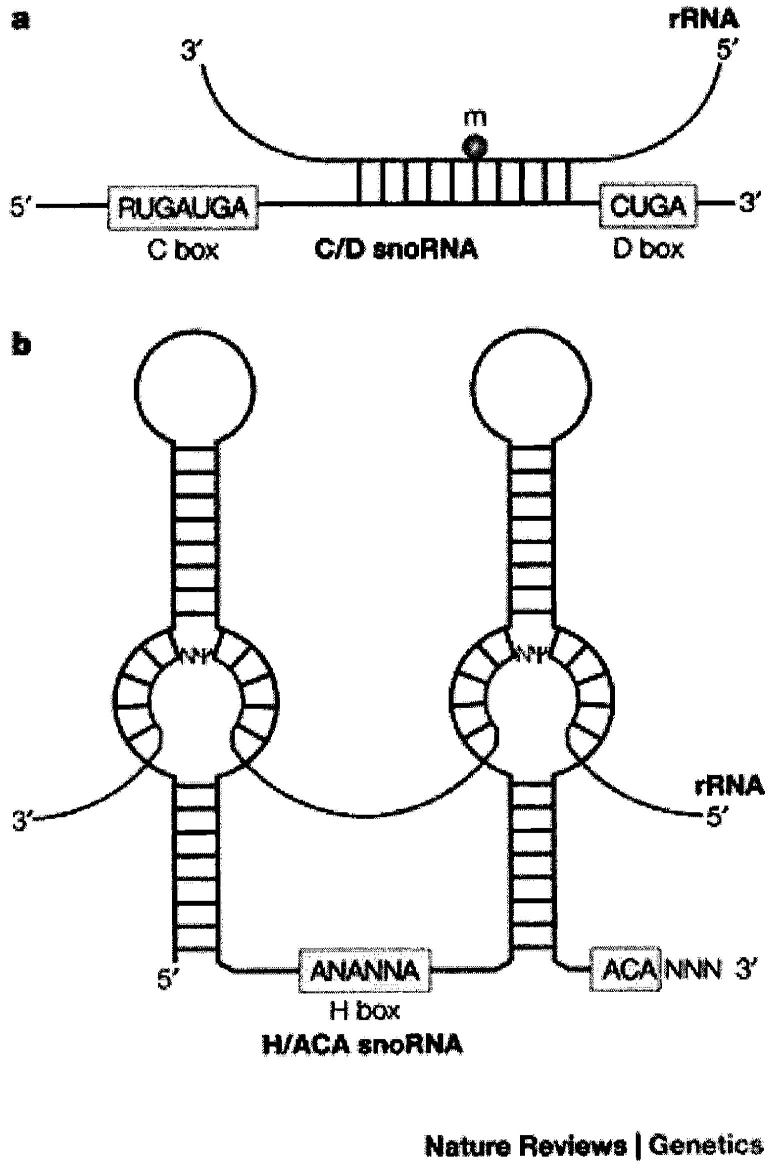

Figure 2-1: snoRNA secondary structure binding rRNA for 2′-O-ribose methylation and pseudouridylation [15]. The ladder-like elements of the diagram represent base pairing.

by manually searching for such complementary mutations since the 1970's [18].

The long-standing approach to automation of RNA secondary structure prediction is estimation of the energetically optimal secondary structure [55, 53]. However, there appear to be some inaccuracies in the thermodynamic models of RNA folding used by these approaches, as secondary structures confirmed by other methods such as comparative analysis are often suboptimal with respect to them. It is likely that tertiary interactions are also influencing the thermodynamics of folding, and RNA tertiary structure seems to be just has hard to predict as protein structure.

There have been some recent attempts to automate the prediction of secondary structure through comparative analysis. The problem of identifying ncRNA orthologs from secondary structure is closely related to prediction of secondary structure in known orthologs. My techniques for ncRNA identification draw on ideas implemented in two programs designed to predict secondary structure: Hofacker's `alifold` [27] and Matthew and Turner's `Dynalign` [36].

The `alifold` program

The `alifold` program predicts RNA secondary structure from an MSA of sequences presumed to be orthologous ncRNA genes. It takes as input an MSA of ncRNA orthologs and augments the thermodynamical model used by the Vienna `RNAfold` program [28] by including scores for the arrangement of complementary nucleotides in each pair of columns in the MSA. For a pair of positions in the sequence i and j, the key element of the score is

$$\Sigma_{X,Y,X',Y'} f_{ij}(XY) f_{ij}(X'Y') d_H(XY, X'Y')$$

where X, Y, X', Y' range over RNA nucleotides such that X base-pairs with Y and X' base-pairs with Y', $f_{ij}(XY)$ is the number of rows in which positions i and j contain X and Y respectively, and $d_H(XY, X'Y')$ is the hamming distance, i.e. the number of positions at which the strings "XY" and "$X'Y'$" differ. The distance factor is probably included to give higher weighting to columns in the MSA that exhibit some variation—columns comprised of just one nucleotide each are not very interesting even if they are complementary to each other, as standard MSA algorithms optimize precisely for uniformity within columns. One corollary of this is that comparative analyses of RNA secondary structure are only appropriate for orthologs exhibiting sufficient nucleotide variation. However, another limitation of `alifold`'s approach is that it relies on the MSA accurately aligning orthologous base pairs in the consensus secondary structure to each other. Since programs such as `ClustalW` generate alignments purely from sequence identity, this means there must be enough homology between the sequences to provide accurate clues about how to align non-identical nucleotides. (See table 2.2 for an example of `ClustalW` failing to align orthologous base pairs correctly.)

The `Dynalign` program

The `Dynalign` program [36] is rather remarkable, in that it requires *no* sequence homology whatsoever. Instead, it takes a pair of sequences assumed to be ncRNA orthologs of sufficient similarity that all orthologous features of their secondary structures are offset from each other by at most some small distance M. It examines the local thermodynamics of pairings between 5-, 6- and 7-tuples within each of the two sequences, and determines from these a secondary structure that minimizes the total free energy of both sequences. However, it is very slow. Matthew and Turner [36] state that `Dynalign`'s run time grows cubically in the length of the shortest sequence it is passed, and that the current implementation is only suitable for sequences of length less than 300 nucleotides or so. They also state that `Dynalign`'s algorithm will not scale well to alignments of more than two sequences.

2.1.3 Detection of ncRNA genes

The thermodynamics of secondary structure provides at least some clues to the folding of a lone RNA molecule, but so far searching for thermodynamically stable RNA sequences has proven to be an inadequate technique for identification of ncRNA genes [46]. The leading available program for detecting ncRNA genes combines thermodynamics with comparative analysis of a pair of orthologous ncRNA genes.

Searching for ncRNA genes with QRNA

Rivas' and Eddy's QRNA [47] tests pairwise alignments of candidate ncRNA orthologs by computing their probabilities with respect to a statistical model RNA, which describes the thermodynamics of ncRNA secondary structure and the concomitant mutation patterns described in section 2.1.2. It flags alignments that look more likely with respect to this model than with respect to a model for alignments of regions coding for protein (denoted by COD and a model for regions of arbitrary DNA (denoted by OTH). The models COD and OTH are relatively simple--COD encodes codon mutation frequencies in the six possible translation frames, and OTH encodes mutation frequencies for single nucleotides. The model RNA is a Stochastic Context Free Grammar [14] encoding the stacking energies of RNA base pairs, and rates of mutation in known orthologous ncRNA base pairs in pairs of orthologous ncRNA sequences. The probability they actually compute is

$$\Sigma_S P(\overline{XY}|s, \texttt{RNA})P(s|\texttt{RNA}).$$

The first factor is the probability of the mutations between $\overline{XY}$, given the base pairing specified by s, and the second is the relative probability of the secondary structure s with respect to the space of all possible secondary structures, given the estimated thermodynamic stability of s.

Rivas and Eddy use the Inside Algorithm [14] to compute the odds of an alignment with respect to the RNA model, and its run time grows as the cube of the length of the alignment. This is reasonable for short alignments such as for snoRNA orthologs, but is prohibitively expensive in searches for longer orthologs (the XIST gene in humans is estimated to be about 19 kB long [29].)

For the same reasons as alifold (see subsection 2.1.2,) QRNA depends critically on the accuracy with which the alignments it is passed match up orthologous base pairs. With orthologous sequences with high nucleotide identity, QRNA is reasonably sensitive because the identical nucleotides give many clues as to how to align the non-identical nucleotides. Its sensitivity degrades when passed poorly aligned orthologs, however, as shown in table 2.4 on page 41. Rivas and Eddy state that QRNA performs most reliably on candidate orthologs with 65% – 85% nucleotide identity.

2.2 Identification of MSA's of ncRNA orthologs

The complementary mutation patterns in ncRNA orthologs are very distinctive, and the Bayesian approach used by QRNA fails to exploit this—an alignment with complementary mutations in appropriate positions conforms to QRNA's statistical model, and gets a good RNA score as a result, but this does not preclude it from also scoring well with respect to the COD and OTH models, even though they provide no explanation for the complementarity. It seems as though a more sensitive approach is to score the statistical significance of the complementary mutations an alignment exhibits. The intrinsic sensitivity of the two methods aside, another advantage to significance-based detection is that it easily generalizes to MSA's of large sets of orthologs, which contain far more information than a pairwise alignment.

2.2.1 Column-based significance estimates

My first attempt at detecting MSA's of ncRNA orthologs was based on estimating the statistical significance of the number of complementary mutations seen in each pair of columns in an MSA. This method proved inaccurate in the context of automated searches, but is worth describing as it is a simple precursor and contains the key idea of the next approach. I estimated significance with respect to the null hypothesis that the nucleotides in each column were drawn from an independent random variable whose odds of emitting a nucleotide R are equal to the proportion of R nucleotides in the column. Thus in table 2.1, the "d" columns contain CGGUACGGG and GCCAUGCCC, so if the random variables the columns are assumed to have been drawn from are N_1 and N_2, then the probabilities of drawing A, C, G or U from N_1 are 1/9, 2/9, 1/9 and 5/9 respectively and the probabilities for N_2 are 1/9, 5/9, 2/9 and 1/9 respectively.

The odds in this model of a particular row having complementary base-pairs is

$$P(N_1 = \mathtt{A})P(N_2 = \mathtt{U}) + P(N_1 = \mathtt{U})P(N_2 = \mathtt{A}) +$$
$$P(N_1 = \mathtt{C})P(N_2 = \mathtt{G}) + P(N_1 = \mathtt{G})P(N_2 = \mathtt{C}) +$$
$$P(N_1 = \mathtt{U})P(N_2 = \mathtt{G}) + P(N_1 = \mathtt{U})P(N_2 = \mathtt{G}),$$

i.e. about 0.45. The odds of seeing nine or more events of this probability in a sample of size nine is 0.45^9, i.e. about 0.0039. This probability is called the *complementarity probability.*

The significance of such pairs by themselves is too low to constitute convincing evidence of orthologous base pairs. Even in the short MSA in table 2.1, there are 990 pairs to be considered, so assuming for simplicity that the pair probabilities are being drawn independently (they are not) the odds of seeing a pair with these odds or less is about $1 - (1 - 0.0039)^{990}$, or about 98%. To compensate for this, the evidence from multiple significant pairs needs to be aggregated. To compute this aggregate, I restrict attention to pairs i and j with a complementarity probability less than 0.05, and compute a rough estimate of the probability of the complementary pairs in a neighborhood of i and j. Specifically, suppose there are l valid pairs of positions nearby (i, j), of the form $(i + k \pm 1, j - k \pm 1)$ for $0 < \|k\| \leq 10$. Denote by $p_1, \ldots, p_l$ the complementarity probabilities for these pairs of columns, and assume they are in increasing order. Then for each $m < l$, I compute

$$\binom{l}{m} p_1 \cdots p_m$$

as an estimate of the odds of independently drawing from the null hypothesis l pairs of columns containing a subset of size m whose complementarity odds q_i satisfy the constraint $q_1 \leq p_1, \ldots, q_m \leq p_m$. I use the minimum over m of this expression as the aggregated probability for the pair and its neighborhood. Finally, as the significance for the complementarity in an MSA, I use the maximum aggregated significance over all pairs of positions.

2.2.2 Tuple-based significance estimates

Since ClustalW tends to only misalign orthologous base pairs by one or two positions, I attempted a new estimate of significance using the number of contiguous reverse complementary tuples in a pair of shifting windows. This version was more tolerant of misalignments, and thus more accurate on automatically generated MSA's.

This significance estimate is with respect to the same per-column null hypothesis as before—that each column is drawn from a nucleotide-valued random variable whose distribution is the same as the proportion of nucleotides in the column. For column positions i and j, with $\|i - j\| > 5$, the complementarity in the windows $(i, i+5)$ and $(j, j+5)$ is examined. For each row in the MSA, I find the longest pair of complementary tuples in these windows, and estimate the odds with respect to the null hypothesis of seeing the resulting distribution of tuple lengths. To compute this estimate, first I estimate the odds of seeing complementary tuples of lengths 3 or more, 4 or more, or 5 by simply adding the odds of seeing complementary tuples of the various lengths at each possible position. The product of these odds, q_{ij}, is taken as an estimate of the probability of the complementarity in windows i and j. This is inaccurate, as while the null-hypothesis is independent with respect to the rows of the MSA, it is also symmetric in them. Thus the product of the rows' odds should be multiplied by the number of ways the rows could be rearranged. However, neglecting this factor results in an effective test, while including it results in unreasonably low significance estimates. It is possible that leaving this factor out compensates for the looseness of the other probability estimates in the calculation.

For example, here are a pair of windows taken from the ClustalW alignments in table 2.2. The variable names in the first row represent the random variables the columns of the MSA are assumed to be drawn from in the null hypothesis. The last column gives the length of the longest reverse-complementary tuple in each row.

X_0	X_1	X_2	X_3	X_4	Y_0	Y_1	Y_2	Y_3	Y_4	Lengths
C	C	C	C	C	G	G	G	G	G	5
g	U	C	G	c	a	C	G	A	-	3
C	U	C	G	c	a	C	G	A	G	4
c	U	U	G	C	g	U	A	A	G	4
c	C	C	A	U	a	G	U	G	G	4
c	u	U	C	U	A	G	G	A	U	3

If X and Y are nucleotide-valued random variables, denote by $X||Y$ the event that a nucleotide is drawn from each, and the resulting pair is complementary. The odds of a reverse-complementary 5-tuple are

$$P(\text{5-tuple}) = \prod_{i=0}^{4} P(X_i||Y_{5-i}),$$

while for a reverse-complementary 4-tuple they are estimated by

$$P(\text{4-tuple}) \leq \Sigma_{k,l=0,1} \prod_{i=0}^{3} P(X_{i+k}||Y_{5-l-i}),$$

and similarly for 3-tuples.

The odds for this degree of complementarity is then estimated by the product of the odds estimates over the rows. In this case, there are two rows with reverse-complementary 3-tuples, three rows with 4-tuples and one row with 5-tuples, so the estimate used is

$$P(\text{3-tuple})^2 P(\text{4-tuple})^3 P(\text{5-tuple}).$$

Once the odds q_{ij} for pairs of windows have been computed in this way, they are combined as follows: an initially empty list L of chosen base pairs within the MSA is kept. For each pair i, j, denote by p_{ij} the product of q_{ij} and the number of positions k, l within the sequence with $||k - l|| \leq ||i - j||$ and whose pairing is consistent with the base pairs in L. This is an estimate of the odds of drawing an MSA with this degree of complementary base pair correlation in some region of length $||i - j||$. The lowest q_{ij} is chosen, and the associated pair i, j is added to L. This process is repeated until no significant pairs remain. Finally the odds for the MSA's total correlation in complementary base pairs is estimated by the product of the chosen q_{ij}'s.

2.2.3 Alignment-based method

In order to detect significant reverse-complementary regions that are not perfectly contiguous, I devised a method based on aligning sequences to their reverse complements.

Searching for helices using Smith-Waterman alignment

To identify helices within an RNA sequence, the sequence is aligned to its reverse complement using the Smith-Waterman algorithm [49]. A pair of nucleotides are flagged as a "match" if they can base-pair in RNA, and as a "mismatch" otherwise. The alignments are scored with a match score of 1, mismatch penalty of -2 and gap penalty of -5. These values are largely *ad hoc*, and bear no resemblance to RNA folding energies. However, it is necessary to use a severe gap penalty to prevent unreasonably long alignments from being chosen.

The significance of an alignment is estimated using a modification of the standard Poisson-approximation estimator for the expected number of alignments of a given score [1]. I first compute the estimator in the case of nucleotides having uniform probability. I simulated this case by randomly generating 3100 pairs of sequences each of length 300 and aligning them as described above. For an alignment of sequences of length m and n, the expected number of pairs with score at least S is estimated by

$$Kmne^{-\lambda S}, \tag{2.1}$$

where K and λ are determined by simulation. The distribution of scores was roughly log-linear in the vicinity of scores 15 and 16, so I estimated K and λ from the numbers of pairs having scores exceeding these values. The values I used were 0.7 for λ and 0.02 for K.

By itself, equation 2.1 is very inaccurate in general, as local variations in nucleotide frequencies can change the probability of randomly matching complementary nucleotides. For instance, when all four nucleotides are equally likely, the odds of drawing two nucleotides that are complementary is $6/16 = 0.375$, but if one nucleotide is drawn from that distribution, while the other is certain to be G, the probability of a complementary pair is 0.5. To compensate for this, I replace λ with

$$\lambda_p = \lambda \frac{\log p}{\log 0.375},$$

where p is the probability of drawing a pair of complementary nucleotides.

To estimate the significance of an alignment $\overline{XY}$ of score S, the value I use for p is the probablity of drawing a pair of complementary nucleotides at random, one from each of the collections of nucleotides in the sequence portions X and Y. If $\overline{XY}$ came from aligning a sequence of length m to a separate sequence of length n, the expected number of alignments of score S is estimated by

$$E(S) = Kmne^{-\lambda_p S}. \tag{2.2}$$

Frequently, $\overline{XY}$ comes from alignment of a strand to its own reverse complement. In that case, I instead use the expected number of alignments in a strand of this length occurring within a portion of length less than that covered by $\overline{XY}$. If $\overline{XY}$ comes from self-alignment of a strand of length l, and the difference between the lowest and highest indices of positions in $\overline{XY}$ is L, then I use the following estimate for the expected number of alignments:

$$E(S) = \frac{KlLe^{-\lambda_p S}}{2}. \tag{2.3}$$

That is, I replace the mn factor, the approximate number of positions X and Y could start at, by $lL/2$, the approximate number of positions X and Y could start at if $\overline{XY}$ spans a length of L or less.

Searching for conserved helices

Given an MSA of candidate ncRNA orthologs of length l, I seek helices that are in roughly the same positions in a statistically significant number of the MSA's rows. I take the row with the most nucleotides, strip it of inserted dashes and align it to its reverse complement as described in the previous subsection. For ease of comparison to other rows, the positions of alignments within the stripped sequence are mapped back to the indices of the corresponding columns within the MSA. Local alignments with scores greater than 7 are examined as potential helices. Helices with scores lower than 7 tend to be very common and have extremely low significance, so ignoring them greatly accelerates the algorithm at a negligible cost to its sensitivity.

Let A be an alignment, with score greater than 7 and expected count $E(A)$. Denote by s_{start}, s_{end} and r_{start}, r_{end} the boundaries of the forward and reverse strands of the alignment, respectively. I seek similarly positioned helices in the other rows of the MSA by aligning the portions of them between positions $\max(0, s_{start} - 10), \min(l, s_{end} + 10)$ and positions $\max(0, r_{start} - 10), \min(l, r_{end} + 10)$. In row i, with best-scoring alignment A_i, its expected count $E(A_i)$ is computed using equation 2.2. The significance of this arrangement of helices is then estimated by

$$-\log E(A) + \Sigma_{E(A_i)<1} - \log E(A_i),$$

which is a rough estimate of the log probability that a tightly clustered set of such helices would be found if the nucleotides in each row of the MSA had been drawn independently from each other. The highest such significance is taken as the significance of the MSA.

2.2.4 Tests of the various methods

Test data

To determine the accuracy of these tests, I ran them on MSA's of RNaseP [7] and SRP [25] orthologs, along with contols that I generated artificially from those MSA's.

Artificially generating data for tests of detection schemes is a perilous business, as one can never be sure that the resulting data contains unrepresentative favorable statistical properties. My first tests used control data drawn directly from the null hypothesis used in the significance estimates, and of course the resulting scores were spectacularly well separated from the scores of the MSA's of genuine orthologs. Specifically, given such an MSA of length L containing n sequences, the null hypothesis described in section 2.2.1 involves nucleotide-valued random variables $N_1, \ldots, N_L$. For each genuine MSA of ncRNA orthologs that I tested on, I constructed a control MSA in which the ith column contained n nucleotides drawn from N_i and then shuffled the columns.

This approach to generating control data was similar to Rivas and Eddy's in [47], in which they shuffle the columns of pairwise alignments of genuine ncRNA orthologs. However, the embellishment of drawing new columns from the N_i's resulted in "MSA's" lacking strong homology between rows, a feature that MSA's of actual homologs inevitably possess. When the tests described above are used on real MSA's, such homology can result in the MSA's being assigned much higher scores than they deserve, because all of my tests assume the rows are independent. For instance, if all rows are identical and exhibit reverse complementarity to which one of my tests assigns a probability p, each row contributes a factor p to the significance estimate, even though having seen the event in one row, it is not very remarkable to also see it in the others.

To compensate for this, I repeated all the tests with strong homology removed from the MSA's. To remove it, I used a sliding window of length 20, and for each row in the MSA, computed its sequence similarity to the previous rows within the window. If in some window a row has more than 85% sequence similarity with some prior row, that portion of the row is masked out with dashes.

Masking strong homology in this way resulted in less spectacular separation from controls. At this point, I concluded that redrawing the columns from the null hypothesis had stacked the deck too far in my favor, and repeated the experiments with controls constructed by *only* shuffling the columns of genuine MSA's.

I also tested on MSA's of known mRNA orthologs from the `EGO` database [33]. Some of these exhibited extremely high significances, presumably because they contain biologically important secondary structure. That problem disappeared when I shuffled the columns of the MSA's. However, the distributions of scores for the resulting MSA's was roughly the same as the distributions for the controls constructed by shuffling the columns of ncRNA MSA's, so they are not displayed here.

Tests of the column-based method

This score seems to be fairly reliable at distinguishing *curated* MSA's of orthologous ncRNA genes. The results are in figures 2-2 through 2-8. They are aggregated by two criteria—nucleotide identity in the MSA's and GC content. Here nucleotide identity in a column of an MSA is defined as the proportion of the most common nucleotide out of all the nucleotides the column contains, and nucleotide identity of an MSA is the average of that over its columns. The GC content of an MSA is simply the proportion of nucleotides in the MSA that are G or C. Aggregation by nucleotide identity is included for comparison with later results, because it will become a significant factor when examining the algorithm's performance on automatically generated MSA's. Aggregation by GC content is included for comparison to Rivas' and Eddy's work (see table 2.3,) who aggregate the results for QRNA in this way because its accuracy seems to be significantly influenced by the GC content of the alignments it is passed.

The column-based significance estimates provide excellent separation between curated MSA's of ncRNA genes and control MSA's. However, this is not a realistic test—this approach to automated ncRNA detection can only be useful if it works with alignments generated automatically by programs like ClustalW [26]. The performance of QRNA is significantly worse on alignments generated with BLASTN than on curated alignments, as table 2.4 shows. Unfortunately, so is the performance of column-based significance estimation, as shown by the overlap between the distributions of ncRNA and control MSA's in figures 2-9 through 2-17 and figures 2-34 and 2-37.

The performance of the column-based significance estimates is significantly worse because ClustalW's alignments are optimized purely for nucleotide identity, and frequently fail to align orthologous base pairs as a result. An example of this failure is shown in table 2.2.

Tests of the tuple-based method

As shown in figures 2-18 through 2-33, this estimate gives almost perfect separation between the SRP and control MSA's drawn directly from the null hypothesis.

.abcde.	.ed.cba.
uCCCCCa	aGG-GGGa
ugUCGcc	aaC-GAau
cCUCGcc	aaC-GAGa
cUCUUGC	GUA-AGAu
acCCAUg	aGU-GGau
cuUCUg-	aGG-Auag
cGGGGCg	aGC-UCC-

...abcde..	.edcba.
u--CCCCCaa	aGGGGGa
....bcd...	..dcb..
---gUCGcca	aaCGA-a
...abcd...	..dcba.
---CUCGcca	aaCGAGa
....abcde.	..edcba
---CUUGCuu	gGUAAGa
....bcde..	..edcb.
aa-cCCAUga	aaGUGGa
.....bcd..	..dcb..
agacuUCUga	aaGGAua
.............abcde..	...............edcba.
ugauuccuaaagcGGGGC-g	acagagcagugaacaGCUCCc

Table 2.2: Misalignment of orthologous base pairs by ClustalW. The first section shows curated alignments of portions of SRP RNA genes that are base paired. The letters in the first row indicate consensus base pairs. The second section shows the corresponding portion of an alignment generated by ClustalW. The letters in the intervening rows are above the corresponding nucleotides in the curated alignment. The base pairs have been shifted significantly in the last row, but are only shifted by one or two positions in the other rows.

However, the performance on more reasonable controls constructed by shuffling the columns of MSA's is much weaker, as shown in figures 2-41 and 2-44. This is probably because strong local homology between rows leads to artificially high significance estimates, as described at the beginning of section 2.2.4. This hypothesis is corroborated by the weak separation of scores for MSA's drawn from the null-hypothesis and genuine MSA's with strong local homology masked out, as shown in figures 2-42 and 2-45.

Tests of the alignment-based method

Figures 2-46 through 2-51 show the performance of the helix-based method. It achieves good separation even on the control MSA's constructed by shuffling columns, as shown in figures 2-47 and 2-50. It does much better on the SRP homologs than on the RNaseP homologs, because the consensus SRP secondary structure has a long helix that looks very significant to the alignment-based method. Suprisingly, this method also achieves poor separation between the MSA's drawn from the null-hypothesis and genuine MSA's with strong local homology masked out.

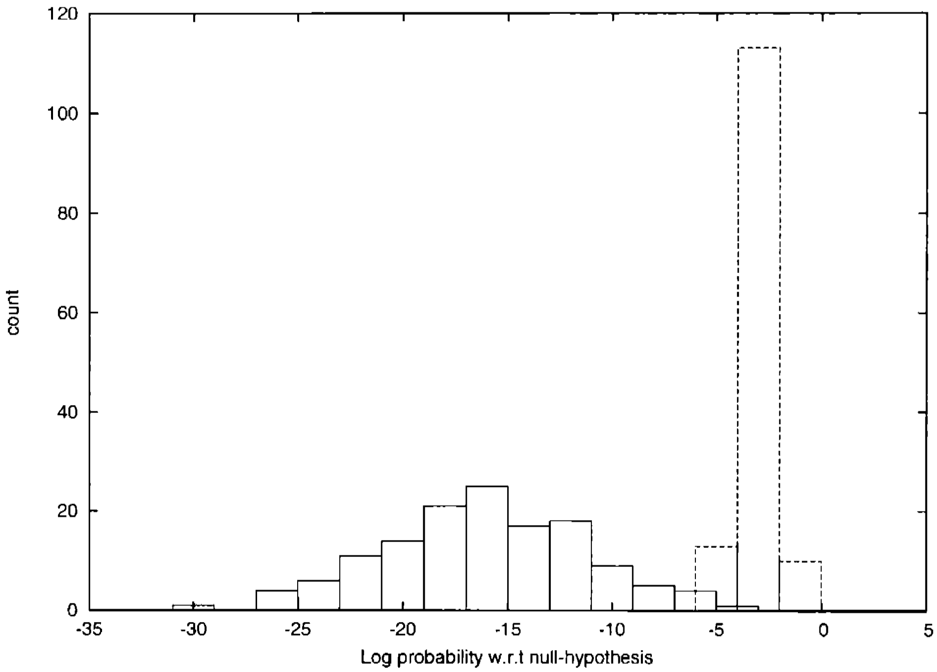

Figure 2-2: Distribution of column-based significance estimates for curated SRP MSA's with 20-30% GC content. Solid bars represent the distribution of significances for genuine MSA's, dashed bars the distribution for control MSA's.

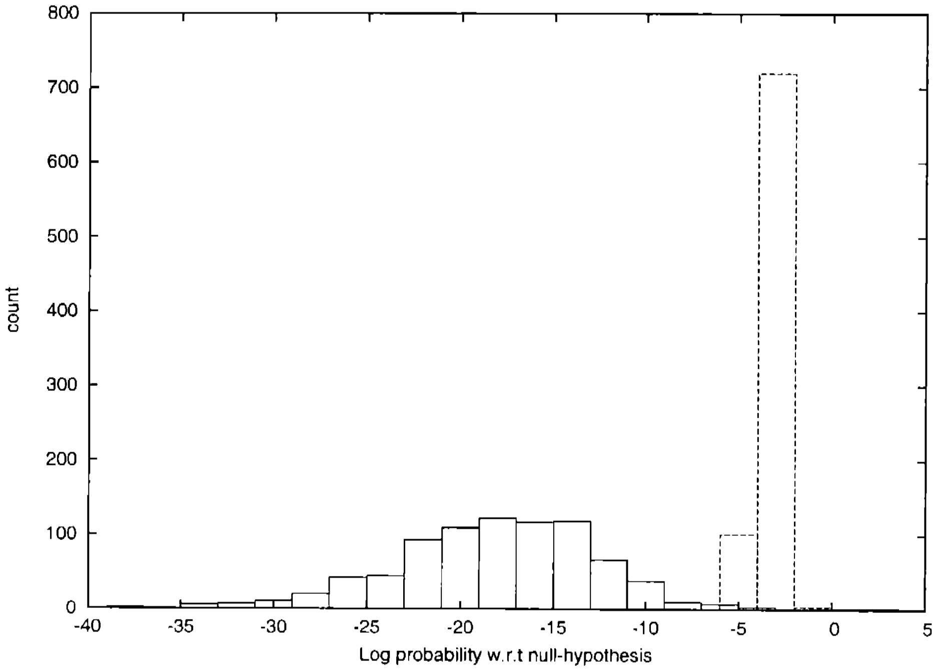

Figure 2-3: Distribution of column-based significance estimates for curated SRP MSA's with 30-40% GC content. See figure 2-2 for more information.

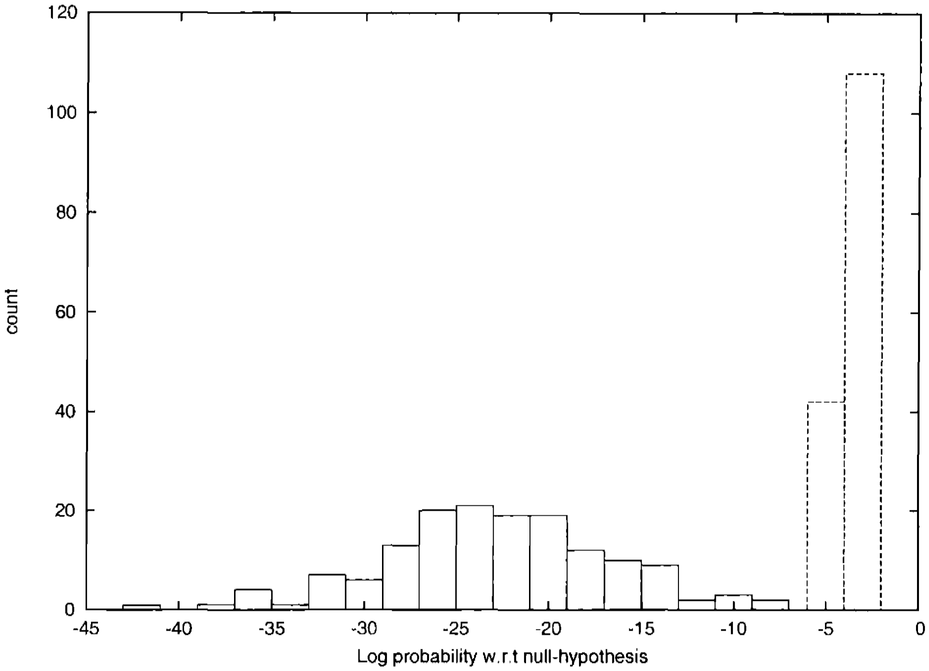

Figure 2-4: Distribution of column-based significance estimates for curated SRP MSA's with 40-50% GC content. See figure 2-2 for more information.

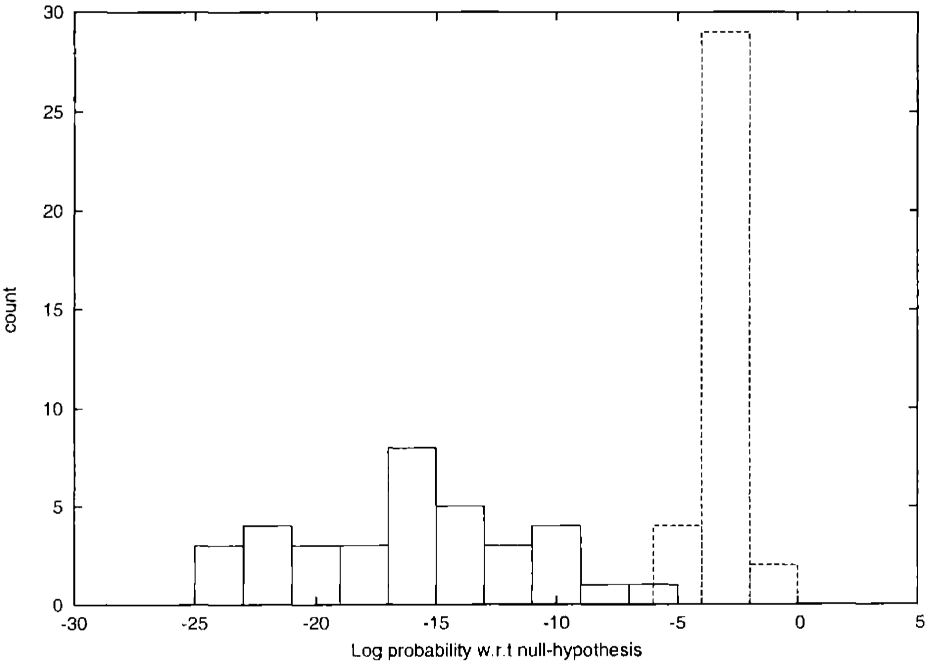

Figure 2-5: Distribution of column-based significance estimates for curated SRP MSA's with 20-30% nucleotide identity. See figure 2-2 for more information.

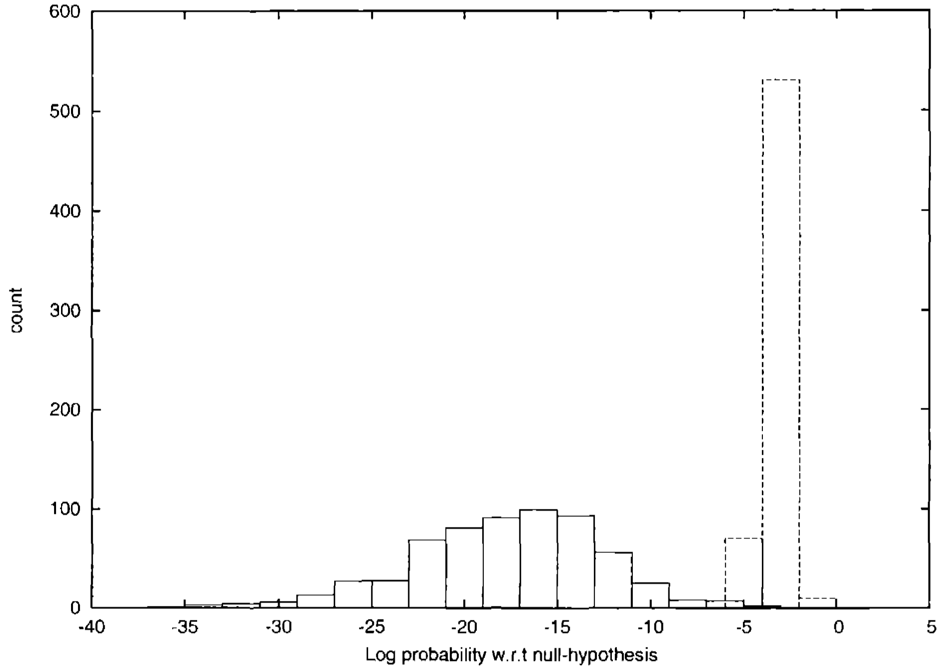

Figure 2-6: Distribution of column-based significance estimates for curated SRP MSA's with 30-40% nucleotide identity. See figure 2-2 for more information.

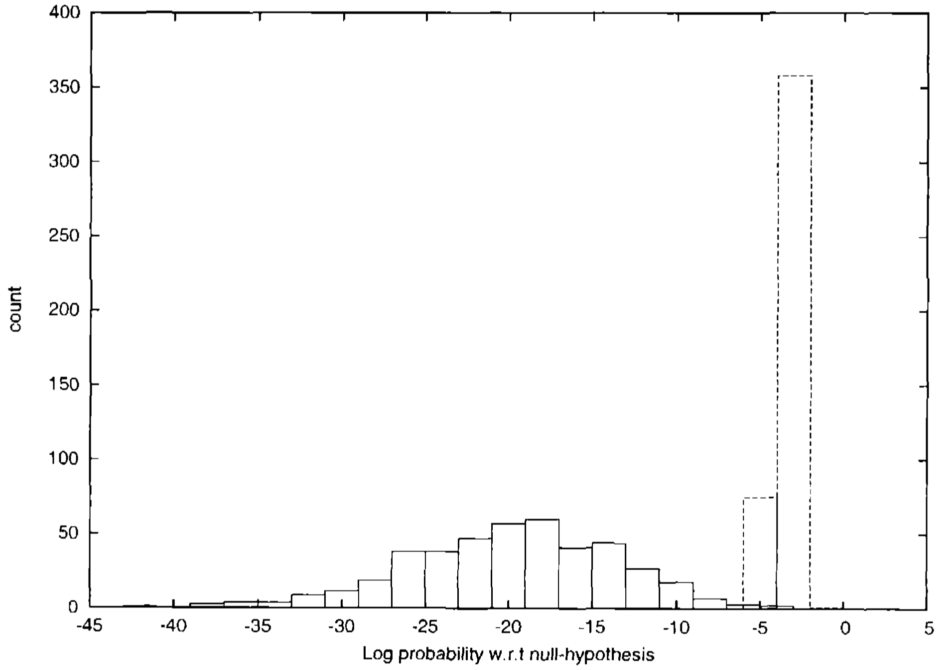

Figure 2-7: Distribution of column-based significance estimates for curated SRP MSA's with 40-50% nucleotide identity. See figure 2-2 for more information.

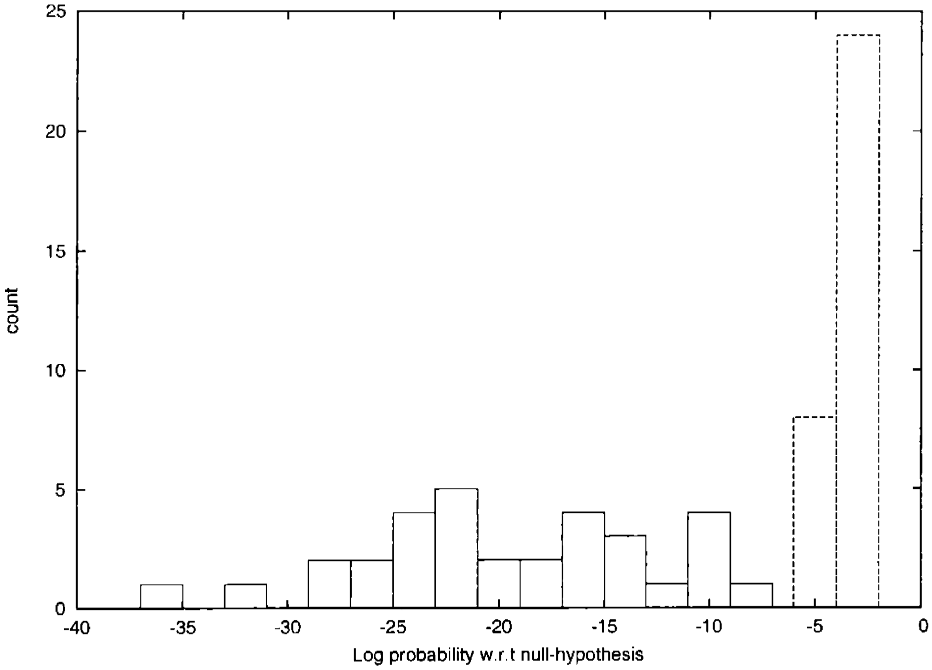

Figure 2-8: Distribution of column-based significance estimates for curated SRP MSA's with 50-60% nucleotide identity. See figure 2-2 for more information.

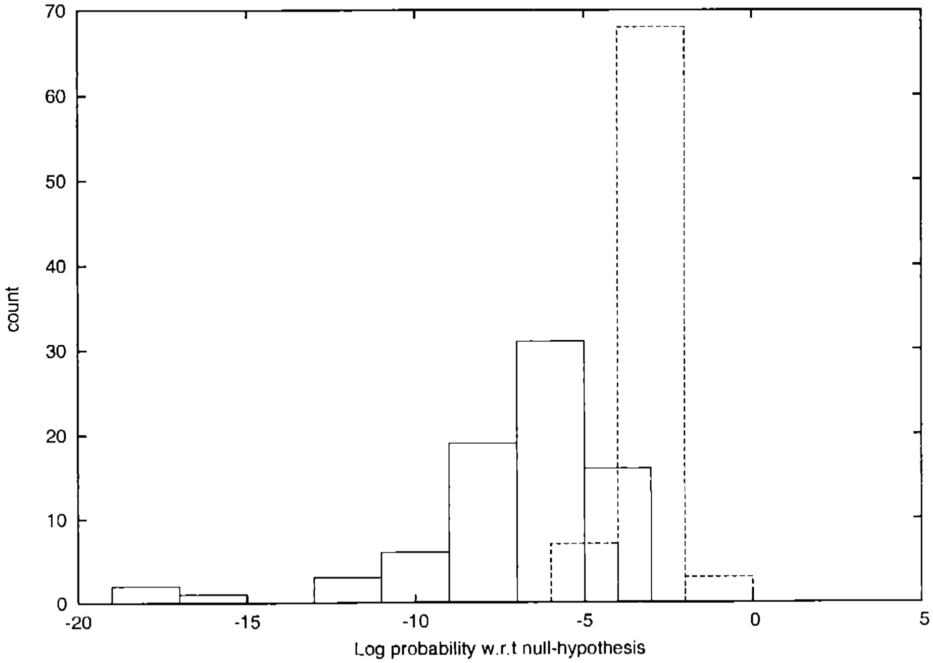

Figure 2-9: Distribution of column-based significance estimates for ClustalW SRP MSA's with 20-30% GC content. See figure 2-2 for more information.

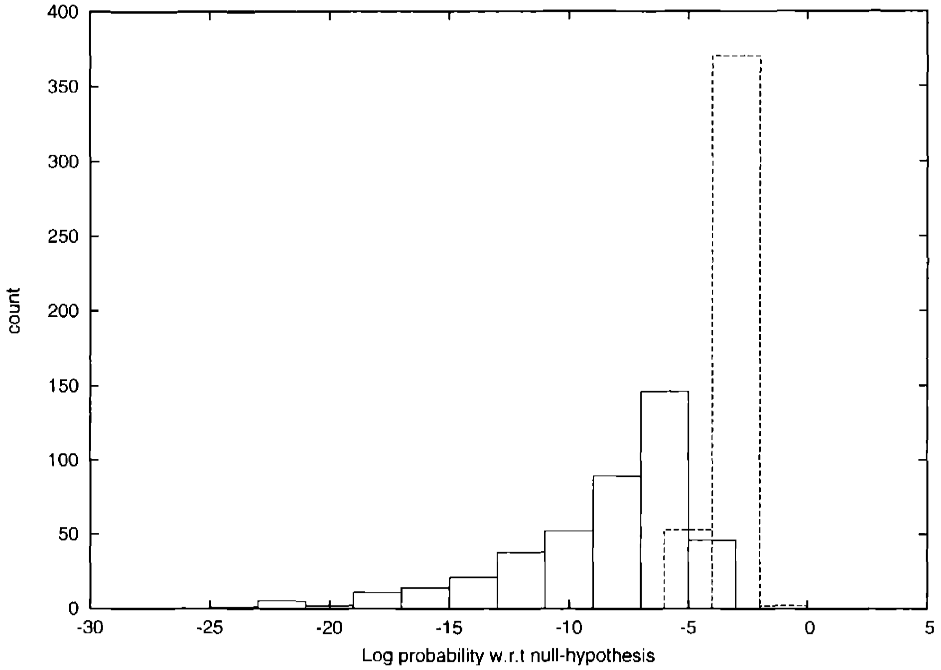

Figure 2-10: Distribution of column-based significance estimates for `ClustalW` SRP MSA's with 30-40% `GC` content. See figure 2-2 for more information.

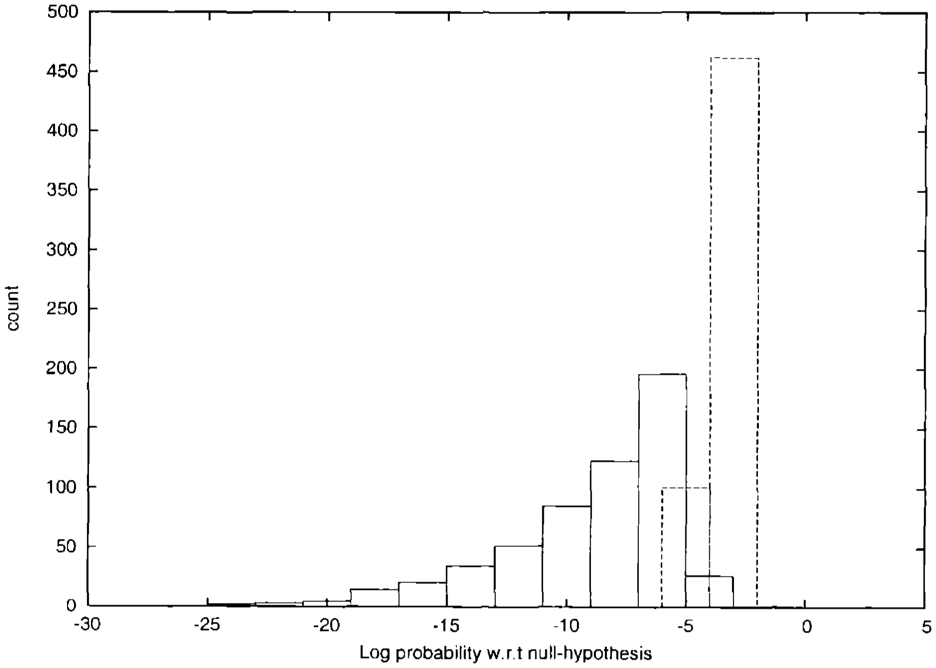

Figure 2-11: Distribution of column-based significance estimates for `ClustalW` SRP MSA's with 40-50% `GC` content. See figure 2-2 for more information.

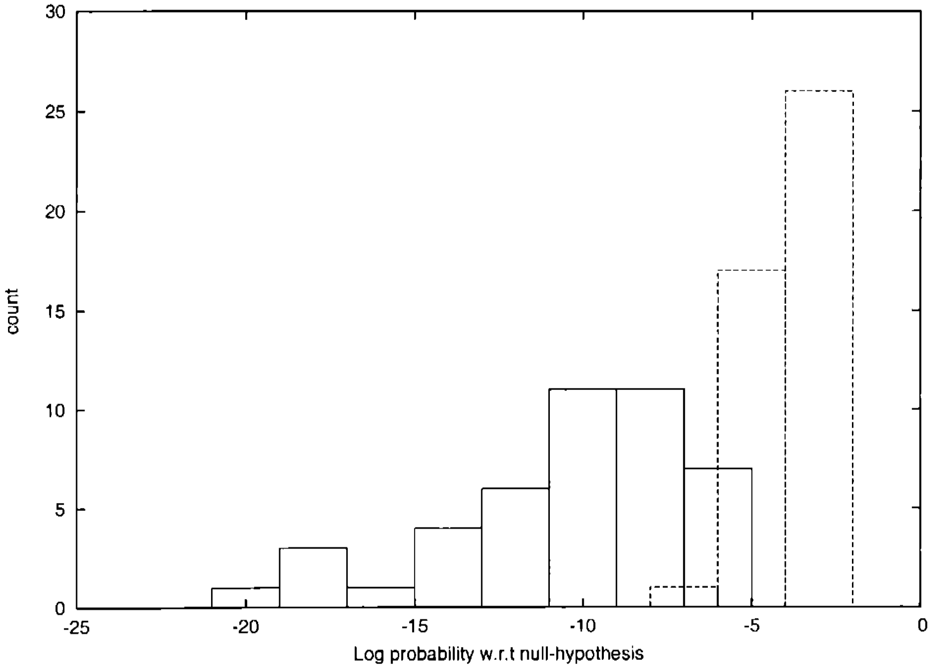

Figure 2-12: Distribution of column-based significance estimates for ClustalW SRP MSA's with 50-60% GC content. See figure 2-2 for more information.

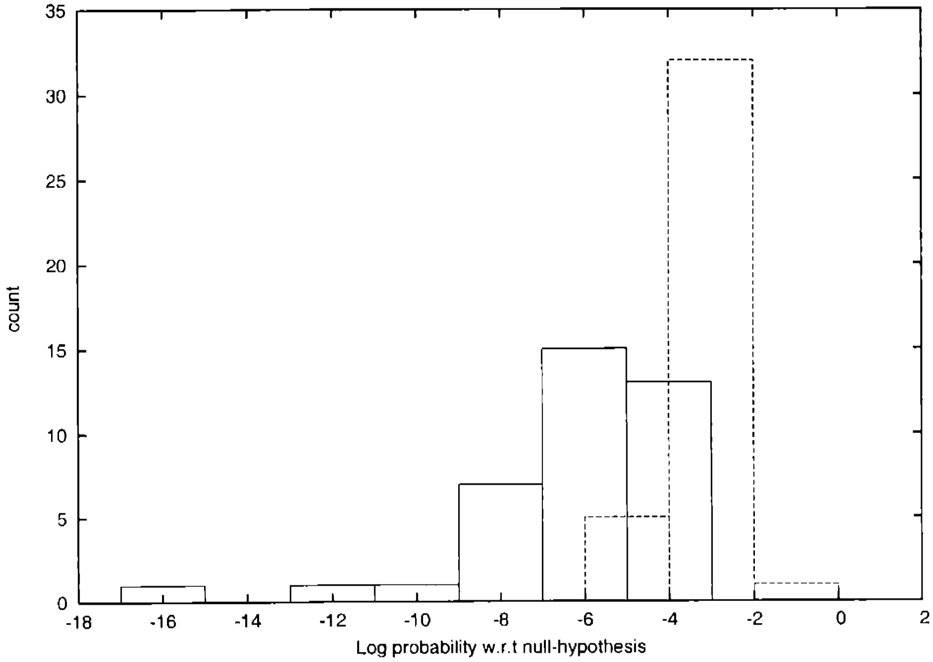

Figure 2-13: Distribution of column-based significance estimates for ClustalW SRP MSA's with 20-30% nucleotide identity. See figure 2-2 for more information.

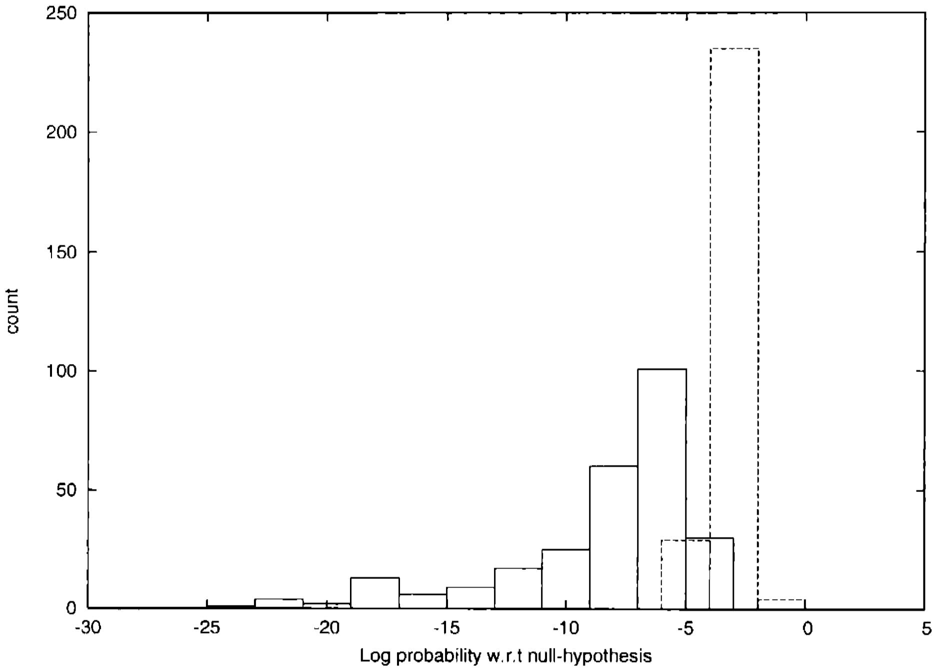

Figure 2-14: Distribution of column-based significance estimates for ClustalW SRP MSA's with 30-40% nucleotide identity. See figure 2-2 for more information.

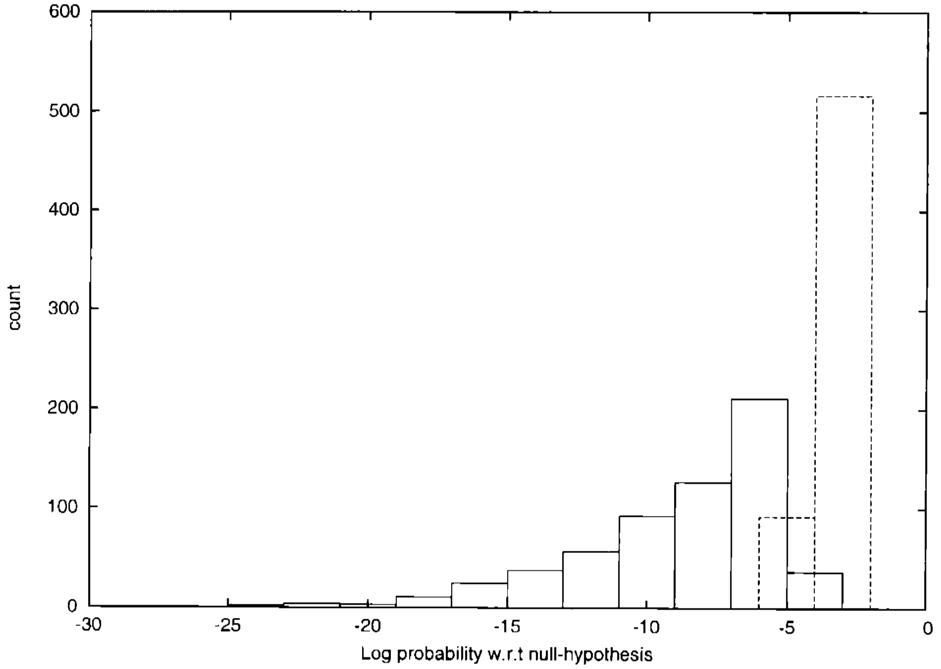

Figure 2-15: Distribution of column-based significance estimates for ClustalW SRP MSA's with 40-50% nucleotide identity. See figure 2-2 for more information.

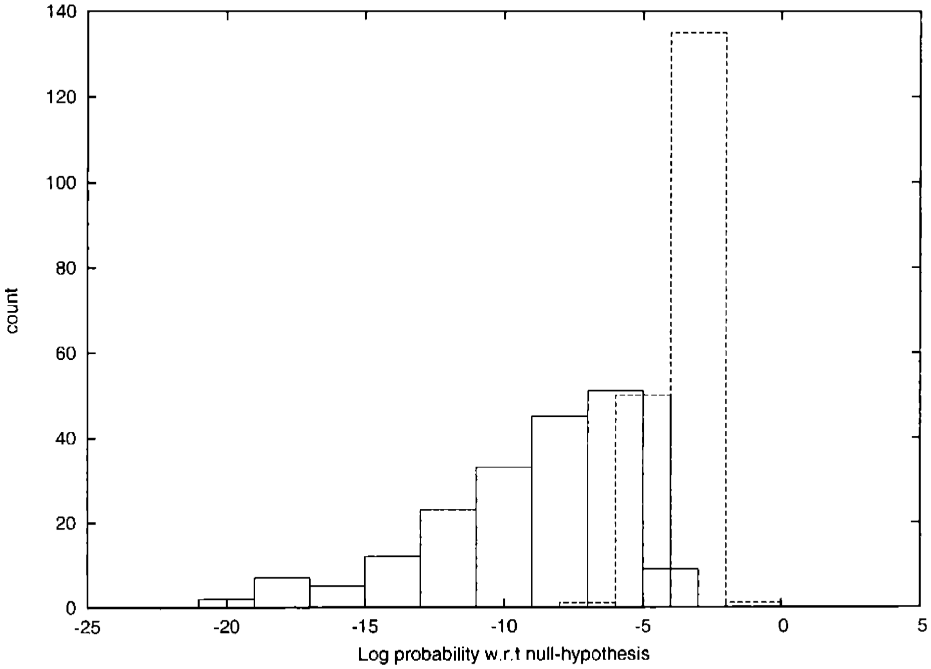

Figure 2-16: Distribution of column-based significance estimates for ClustalW SRP MSA's with 50-60% nucleotide identity. See figure 2-2 for more information.

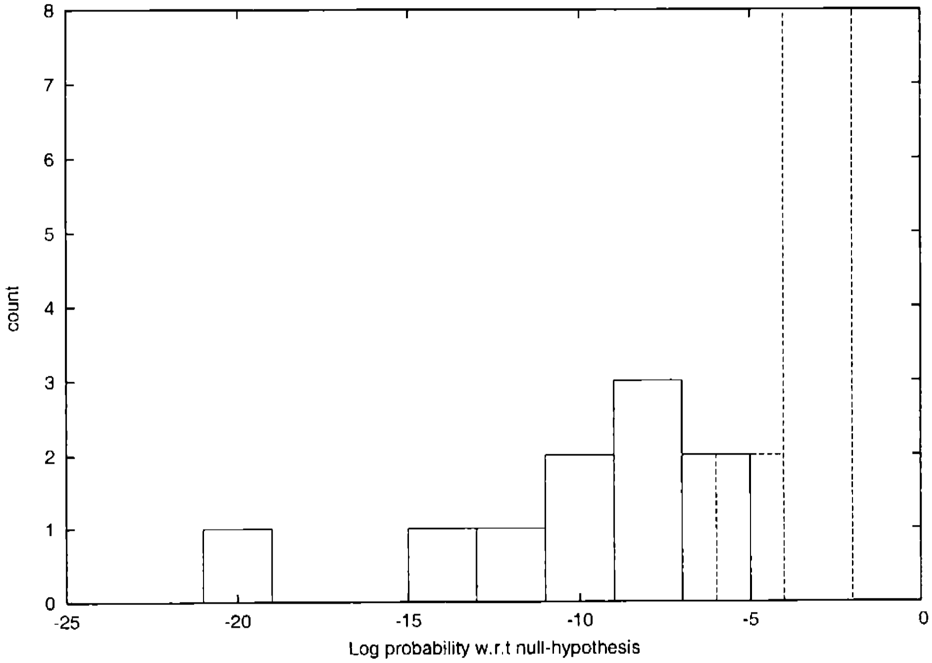

Figure 2-17: Distribution of column-based significance estimates for ClustalW SRP MSA's with 60-70% nucleotide identity. See figure 2-2 for more information.

% Nucleotide identity	# alignments	% sensitivity	% Specificity
0 < 10	140	42.8 (60)	100.0 (0)
10 < 20	827	59.6 (493)	100.0 (0)
20 < 30	503	71.4 (359)	100.0 (0)
30 < 40	764	75.1 (574)	100.0 (0)
40 < 50	283	58.6 (166)	100.0 (0)
50 < 60	434	81.3 (353)	100.0 (0)
60 < 70	88	80.7 (71)	100.0 (0)
70 < 80	70	91.4 (64)	97.1 (2)
80 < 90	73	97.3 (71)	79.4 (15)
90 < 100	61	93.4 (57)	27.9 (44)
100	99	93.9 (93)	29.3 (70)
% GC content			
35 < 40	31	51.6 (16)	93.5 (2)
40 < 45	343	69.1 (237)	96.5 (12)
45 < 50	1131	72.4 (819)	97.9 (24)
50 < 55	1320	69.2 (914)	96.5 (46)
55 < 60	508	73.0 (371)	91.3 (44)
60 < 65	9	44.4 (4)	66.7 (3)

Table 2.3: Sensitivity and specificity of QRNA on curated alignments of RNaseP and SRP orthologs with varying levels of nucleotide identity and GC content. The numbers in brackets in the sensitivity and specificity columns are respectively the number of accurate and inaccurate detections of ncRNA orthologs. Control alignments were generated by shuffling the order of the columns in alignments of genuine orthologs [47].

% Nucleotide identity	# alignments	% sensitivity	% Specificity
60 < 70	419	15.3(64)	99.5(2)
70 < 80	269	26.8(72)	98.5(4)
80 < 90	131	61.1(80)	89.5(19)
90 < 100	78	97.4(76)	67.9(53)
100	106	92.4(98)	24.5(80)
% GC content			
35 < 40	30	6.6 (2)	100.0 (0)
40 < 45	98	40.8 (40)	89.8 (10)
45 < 50	278	39.6 (110)	89.2 (30)
50 < 55	359	35.4 (127)	88.3 (42)
55 < 60	218	46.8 (102)	76.1 (52)
60 < 65	17	29.4 (5)	82.3 (3)

Table 2.4: Sensitivity and specificity of QRNA on BLASTN alignments of RNaseP and SRP orthologs with varying levels of nucleotide identity and GC content. See table 2.3 for more information [47].

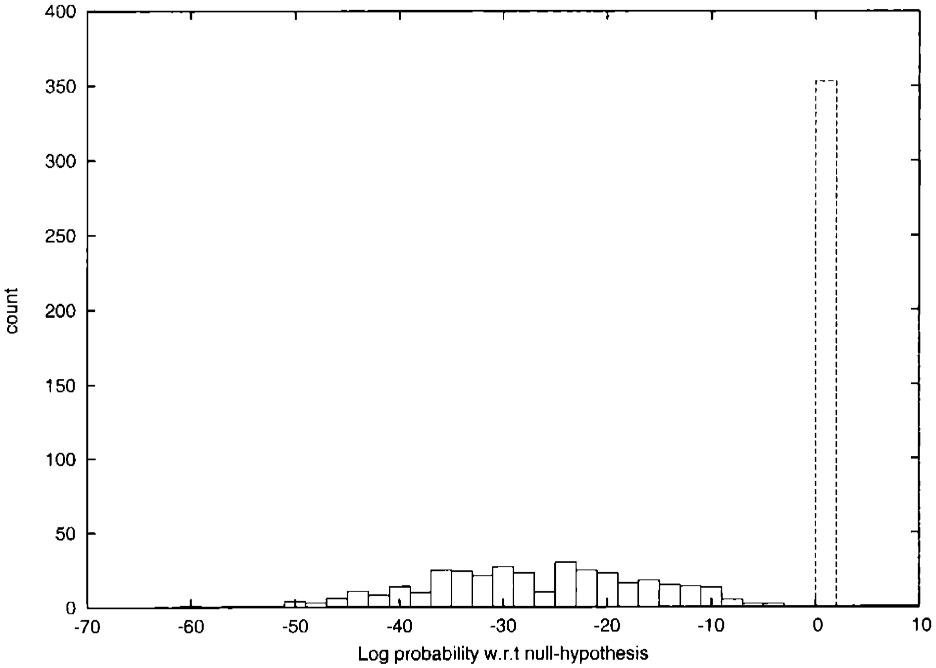

Figure 2-18: Distribution of tuple-based significance estimates for curated SRP MSA's with 20-30% GC content. Solid bars represent the distribution of significances for genuine MSA's, dashed bars the distribution for control MSA's.

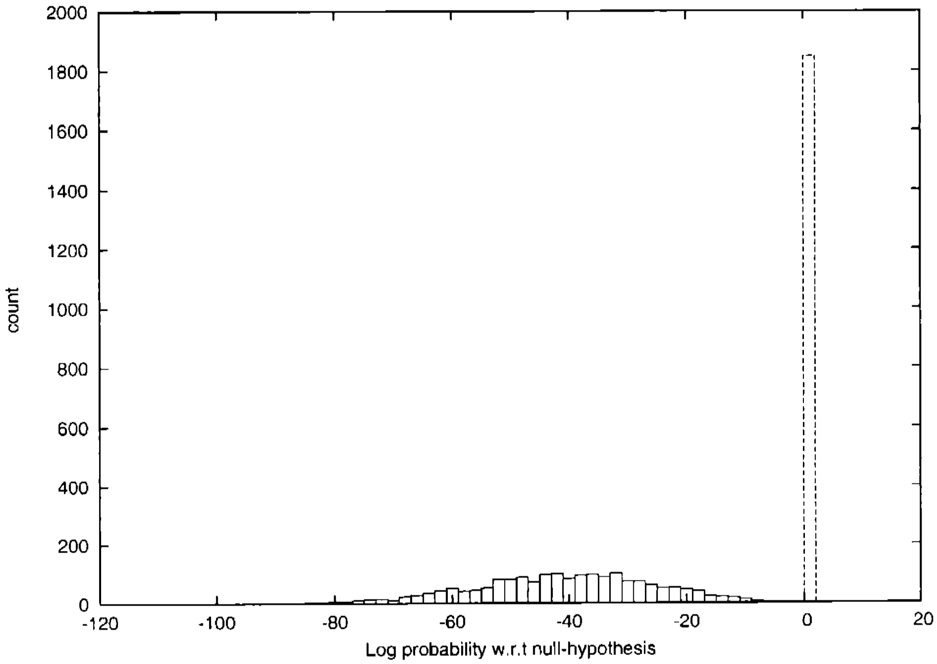

Figure 2-19: Distribution of tuple-based significance estimates for curated SRP MSA's with 30-40% GC content. See figure 2-2 for more information.

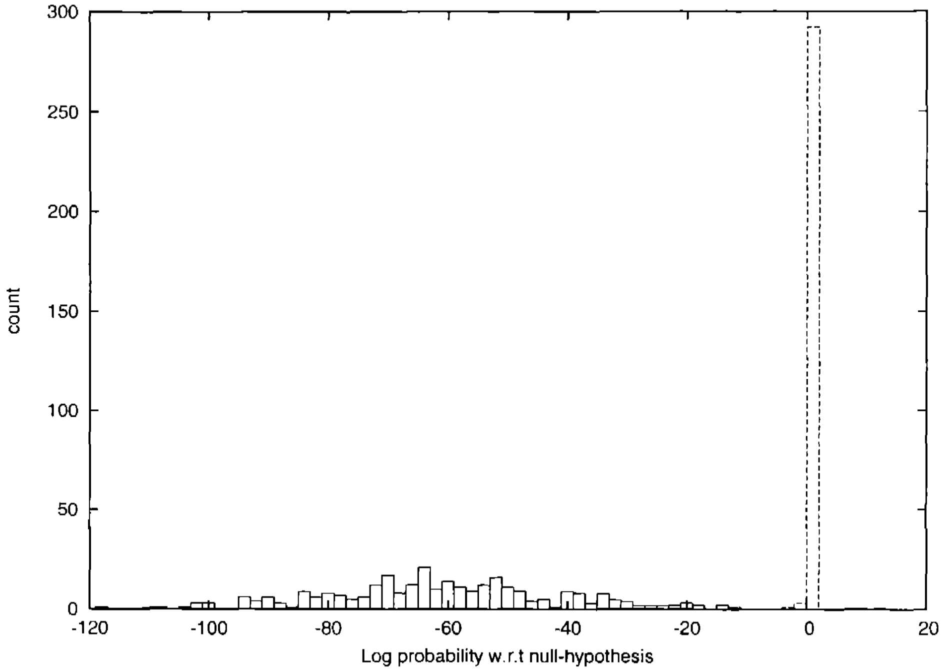

Figure 2-20: Distribution of tuple-based significance estimates for curated SRP MSA's with 40-50% GC content. See figure 2-2 for more information.

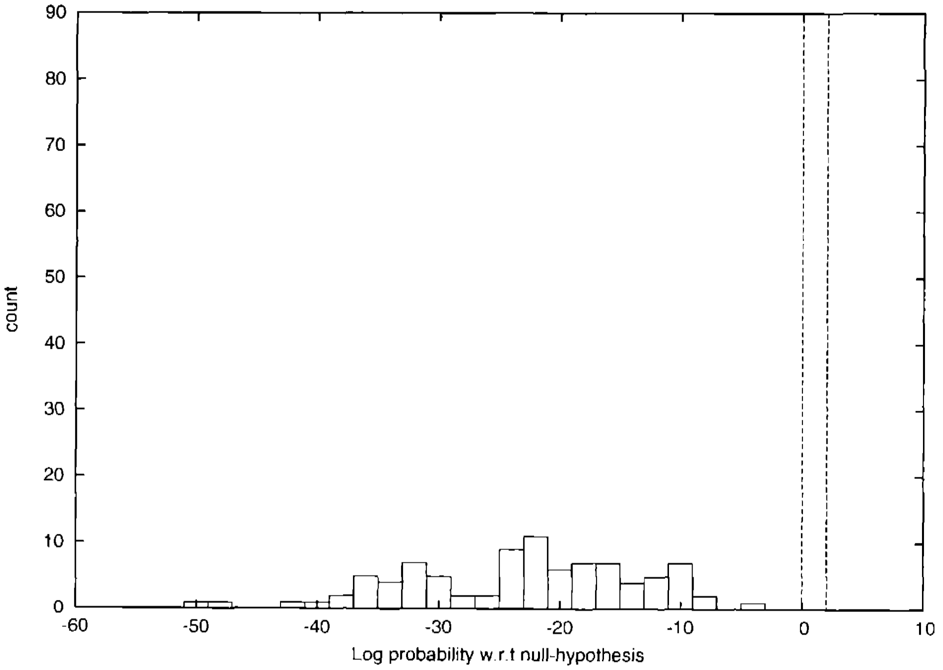

Figure 2-21: Distribution of tuple-based significance estimates for curated SRP MSA's with 20-30% nucleotide identity. See figure 2-2 for more information.

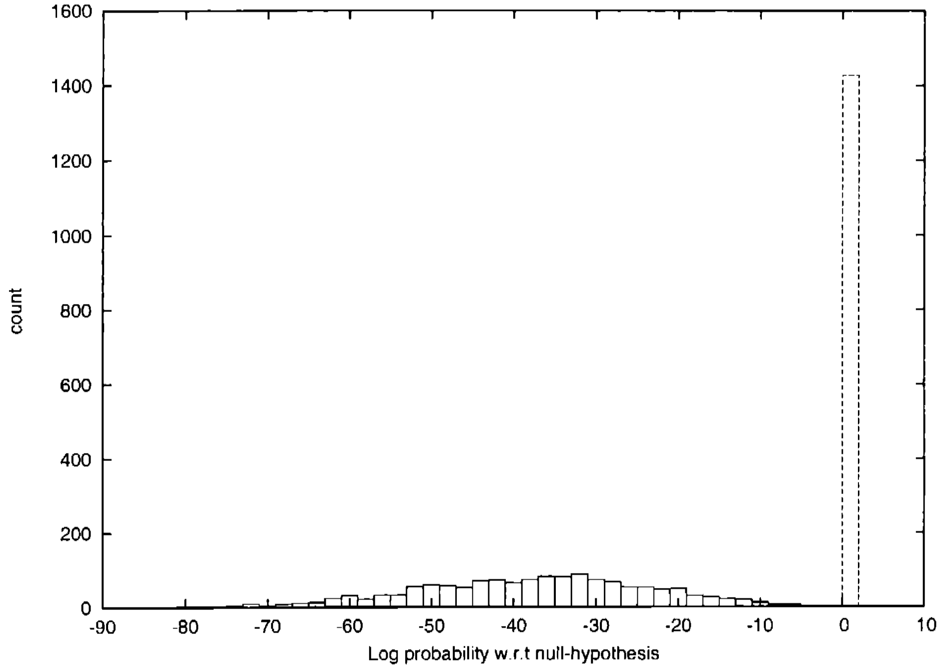

Figure 2-22: Distribution of tuple-based significance estimates for curated SRP MSA's with 30-40% nucleotide identity. See figure 2-2 for more information.

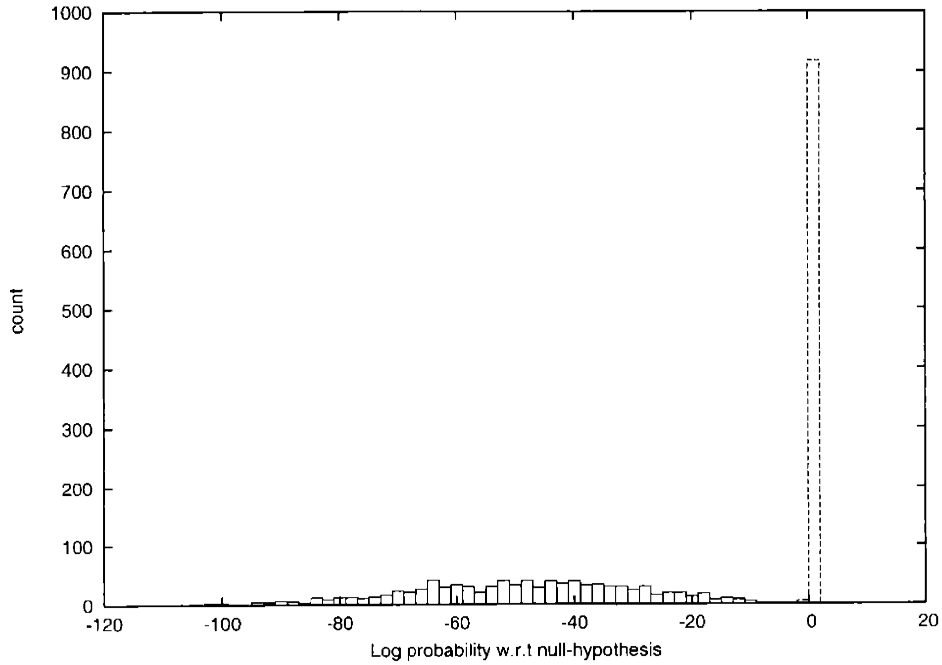

Figure 2-23: Distribution of tuple-based significance estimates for curated SRP MSA's with 40-50% nucleotide identity. See figure 2-2 for more information.

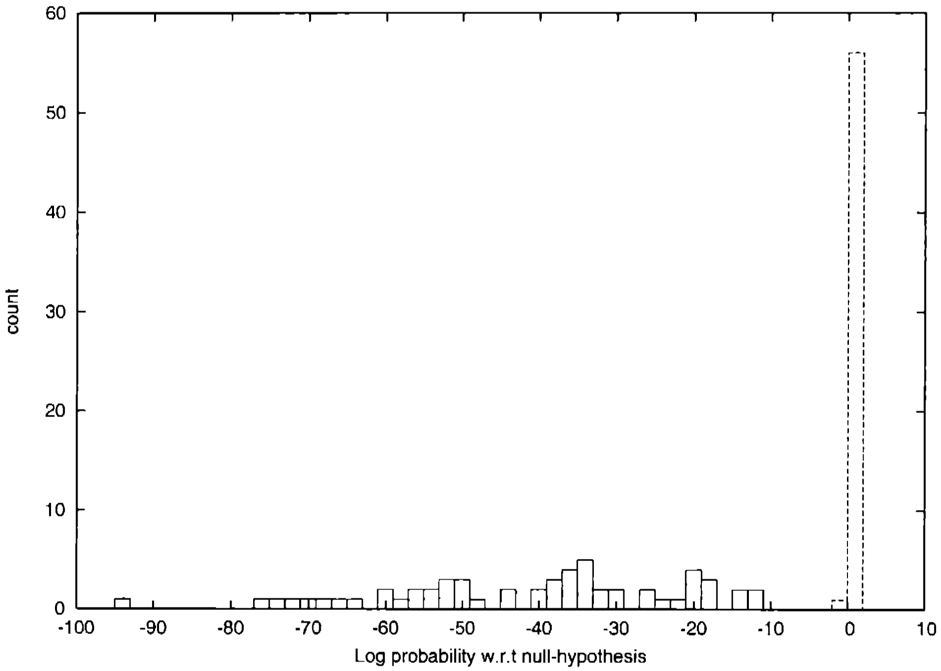

Figure 2-24: Distribution of tuple-based significance estimates for curated SRP MSA's with 50-60% nucleotide identity. See figure 2-2 for more information.

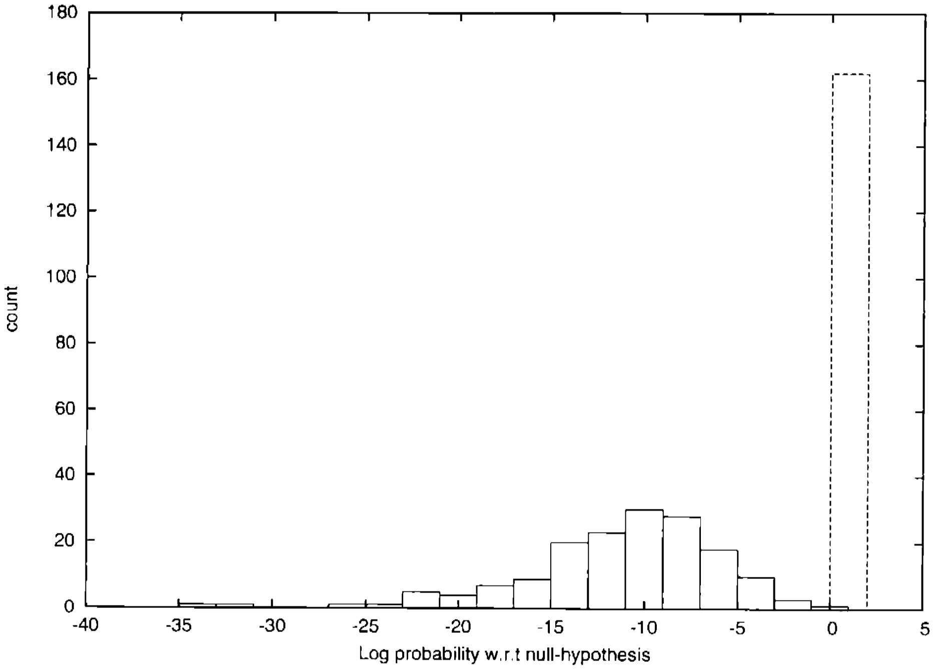

Figure 2-25: Distribution of tuple-based significance estimates for ClustalW SRP MSA's with 20-30% GC content. See figure 2-2 for more information.

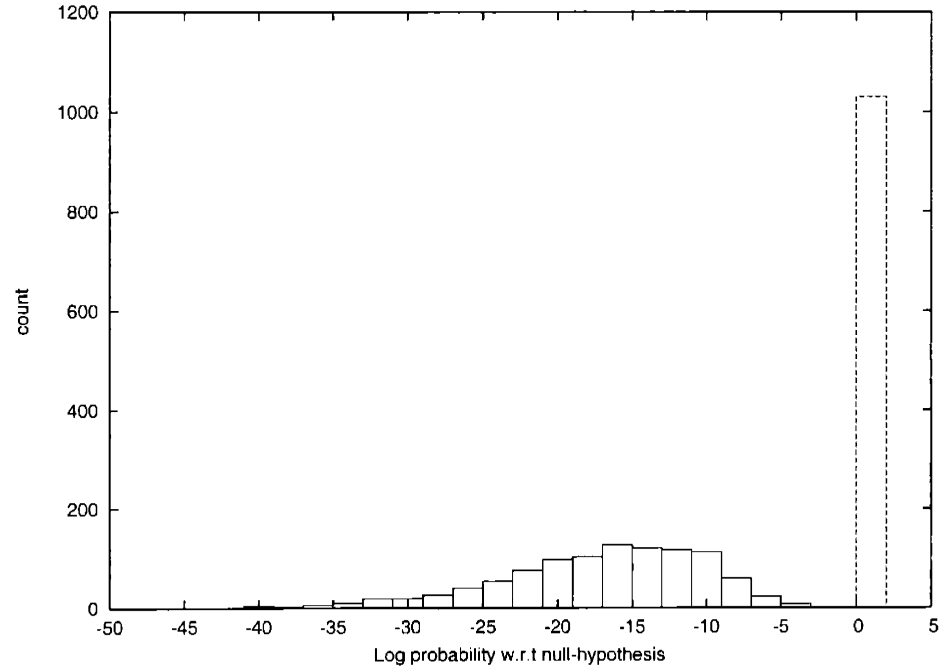

Figure 2-26: Distribution of tuple-based significance estimates for ClustalW SRP MSA's with 30-40% GC content. See figure 2-2 for more information.

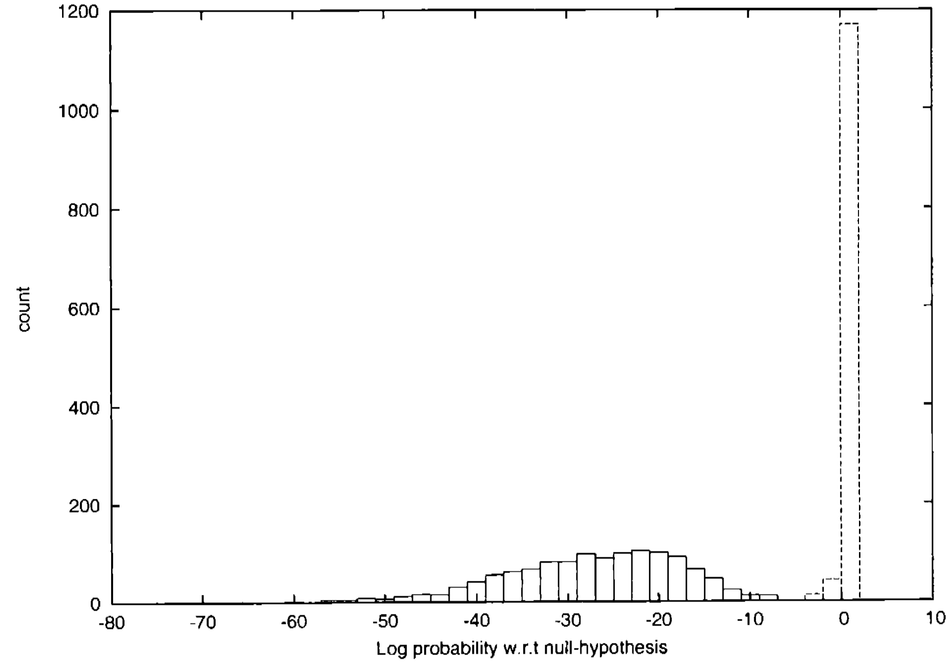

Figure 2-27: Distribution of tuple-based significance estimates for ClustalW SRP MSA's with 40-50% GC content. See figure 2-2 for more information.

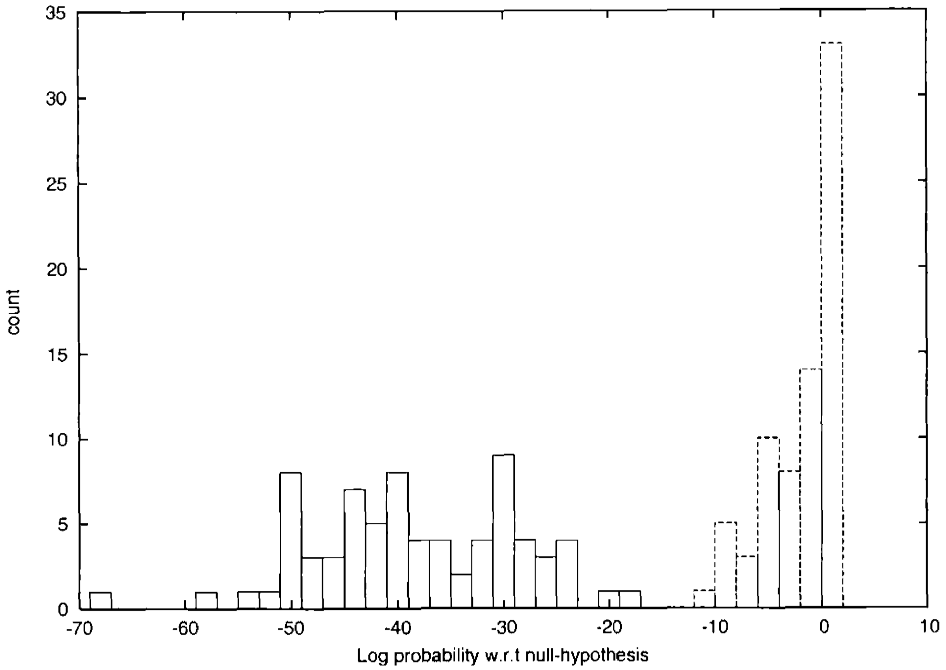

Figure 2-28: Distribution of tuple-based significance estimates for ClustalW SRP MSA's with 50-60% GC content. See figure 2-2 for more information.

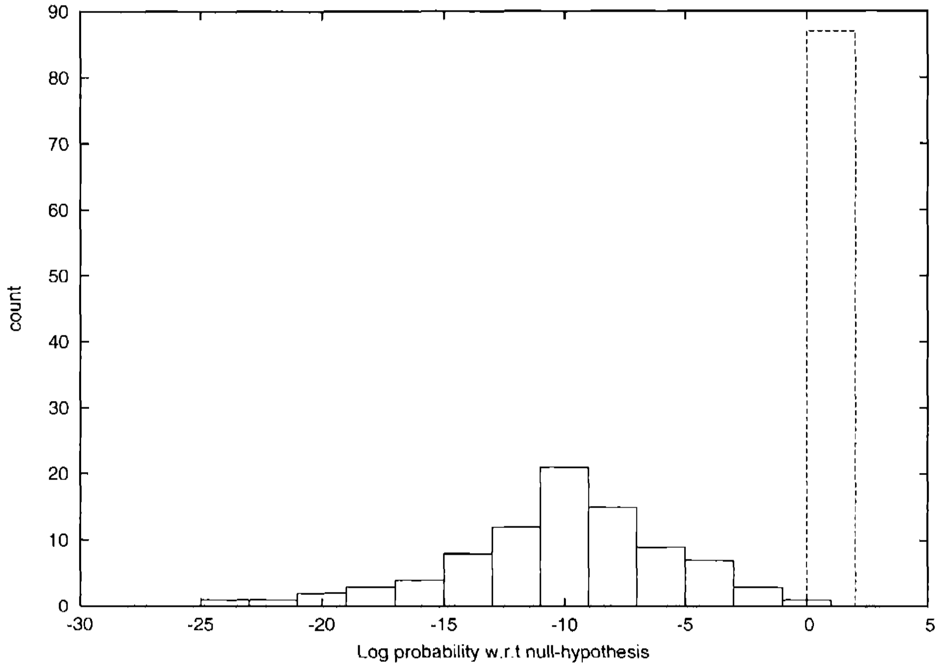

Figure 2-29: Distribution of tuple-based significance estimates for ClustalW SRP MSA's with 20-30% nucleotide identity. See figure 2-2 for more information.

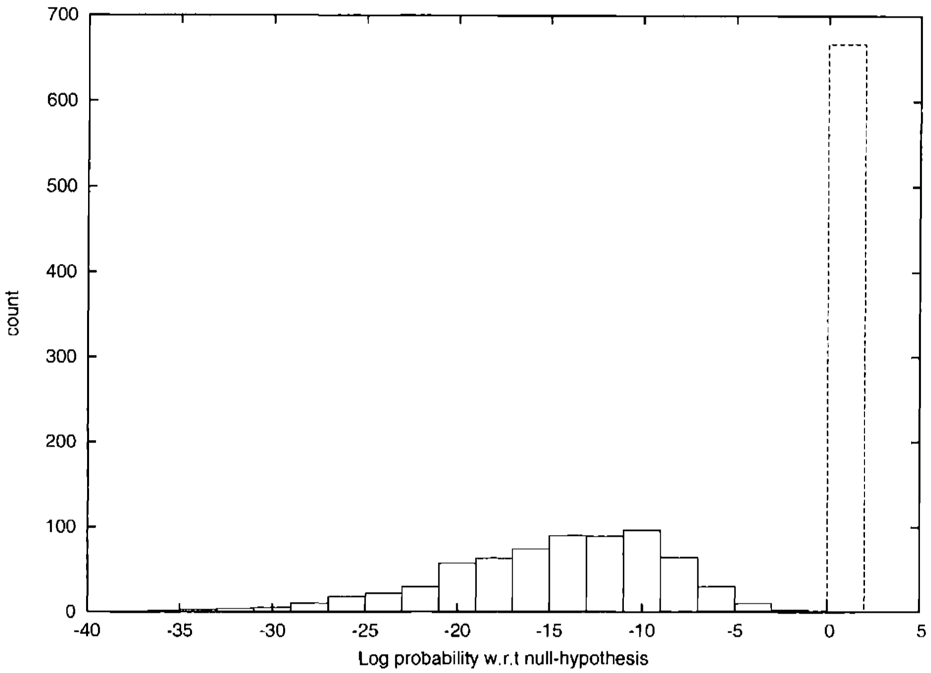

Figure 2-30: Distribution of tuple-based significance estimates for `ClustalW` SRP MSA's with 30-40% nucleotide identity. See figure 2-2 for more information.

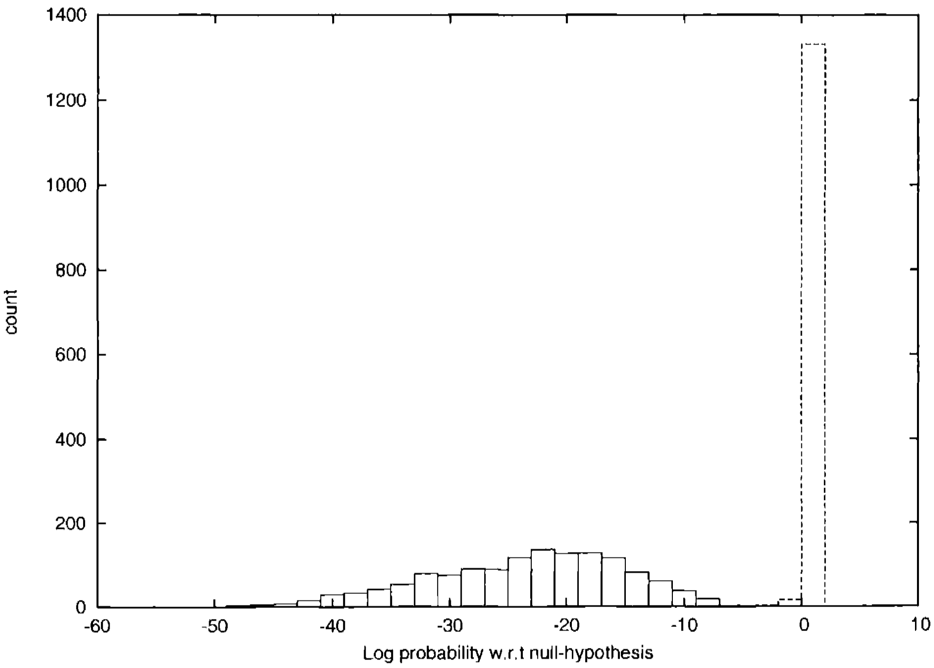

Figure 2-31: Distribution of tuple-based significance estimates for `ClustalW` SRP MSA's with 40-50% nucleotide identity. See figure 2-2 for more information.

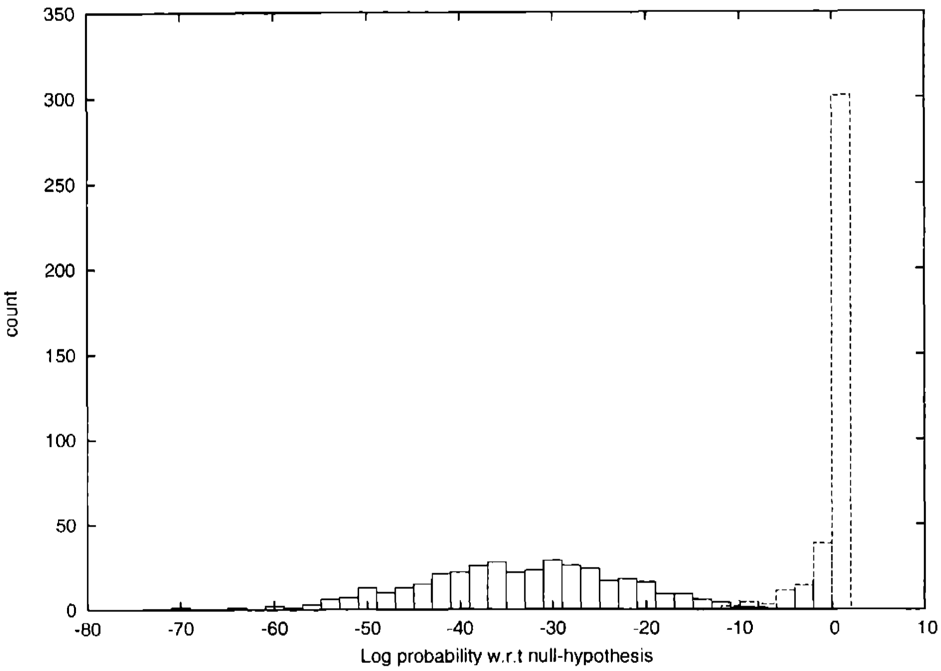

Figure 2-32: Distribution of tuple-based significance estimates for ClustalW SRP MSA's with 50-60% nucleotide identity. See figure 2-2 for more information.

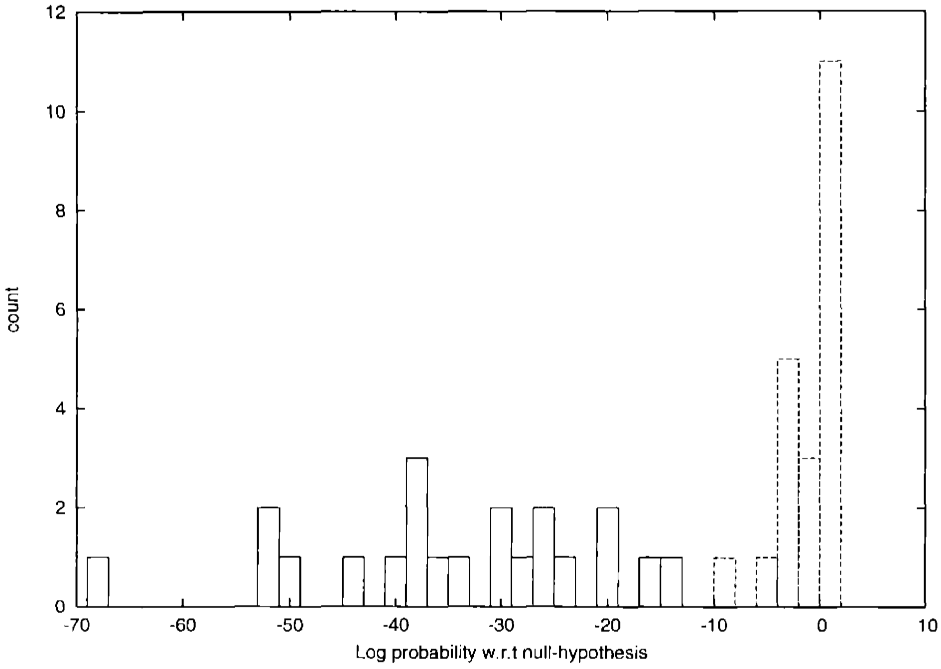

Figure 2-33: Distribution of tuple-based significance estimates for ClustalW SRP MSA's with 60-70% nucleotide identity. See figure 2-2 for more information.

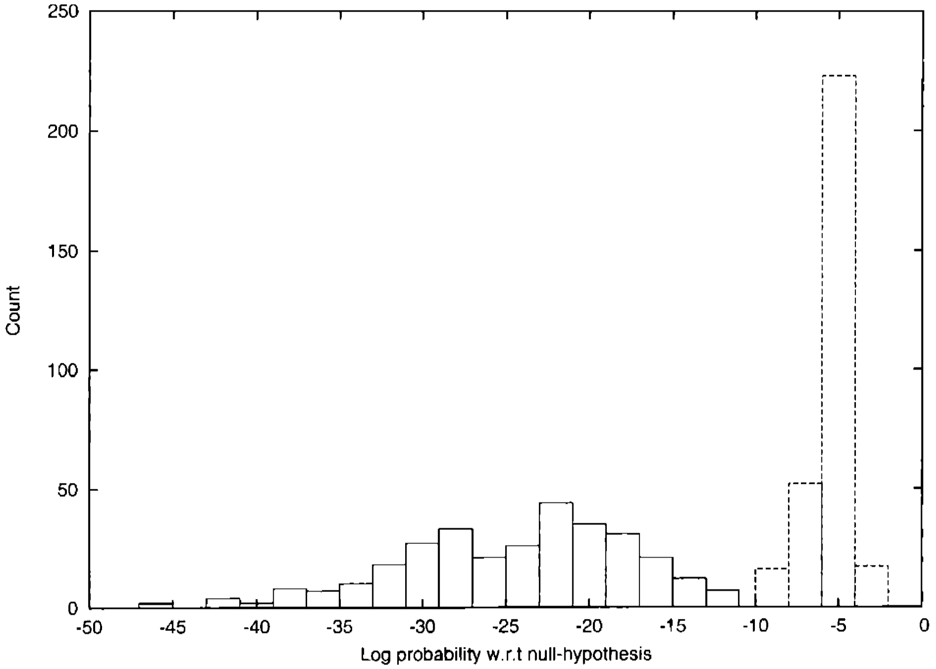

Figure 2-34: Distribution of column-based significance estimates for `ClustalW` RNaseP MSA's versus "MSA's" drawn from the null hypothesis.

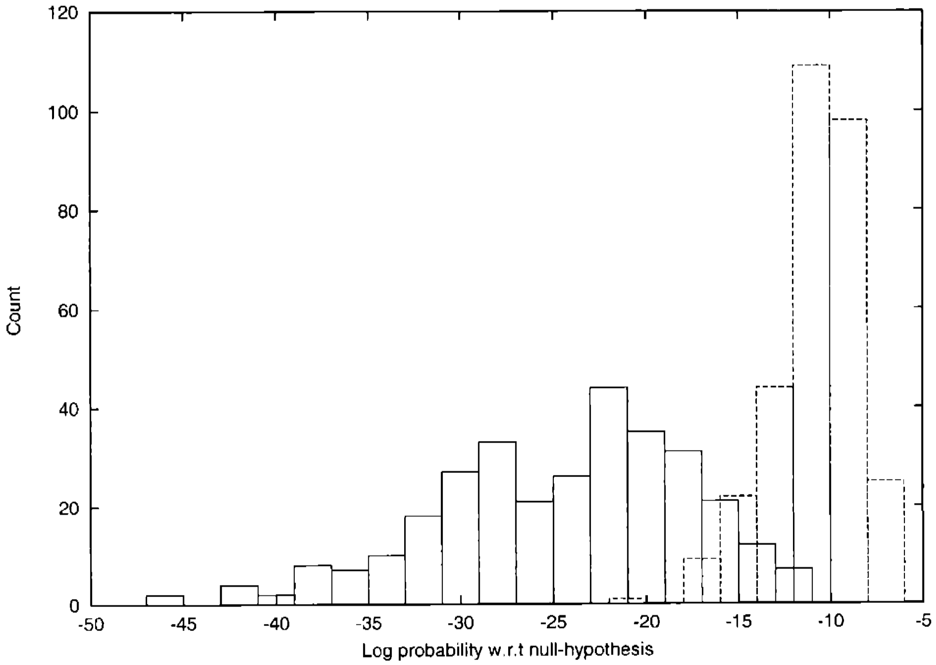

Figure 2-35: Distribution of column-based significance estimates for `ClustalW` RNaseP MSA's versus "MSA's" constructed by shuffling the columns of genuine MSA's.

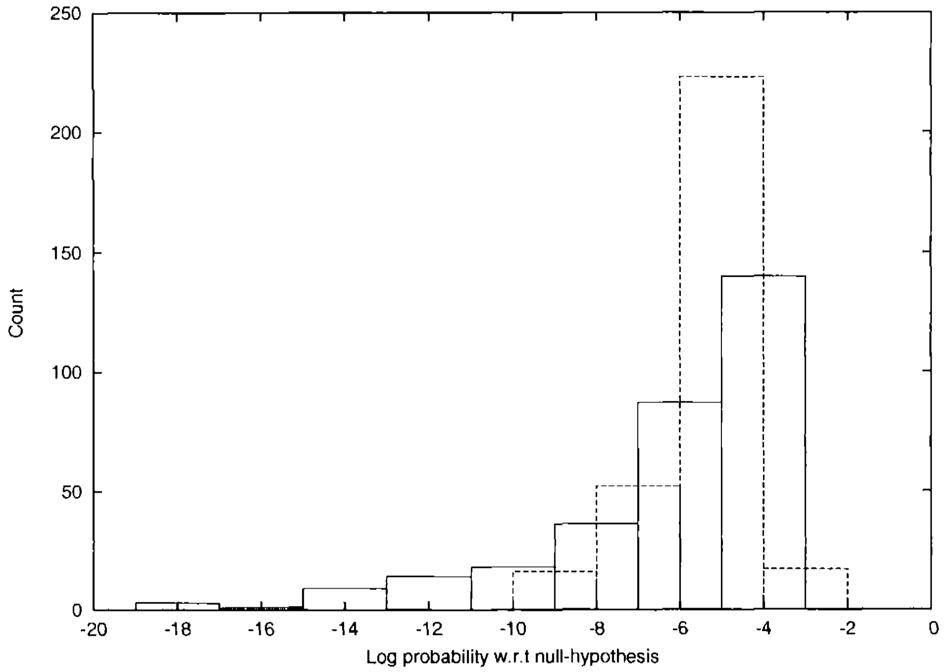

Figure 2-36: Distribution of column-based significance estimates for ClustalW RNaseP MSA's with strong local inter-row homology masked out versus "MSA's" drawn from the null hypothesis.

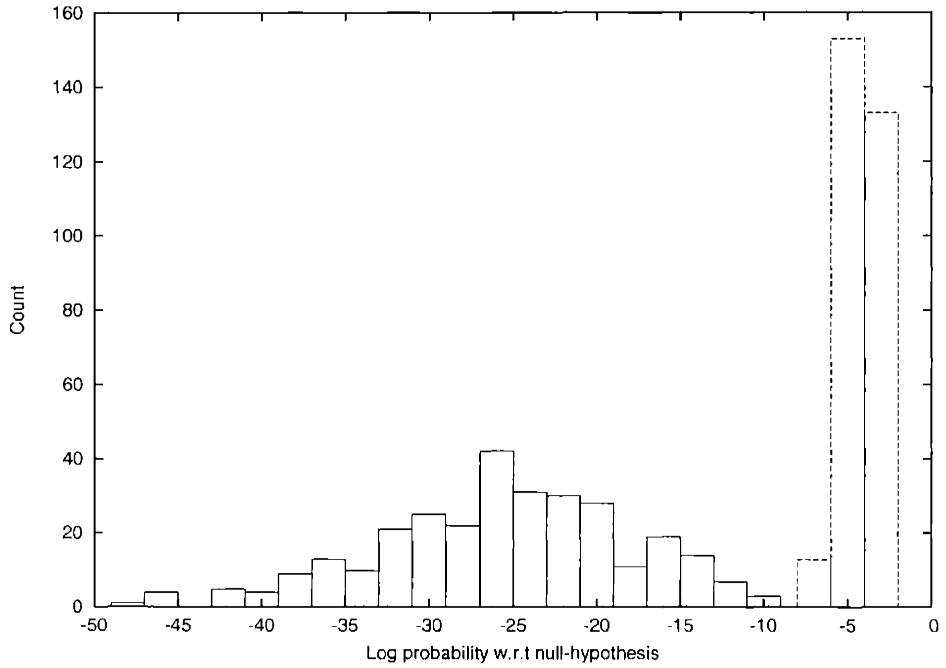

Figure 2-37: Distribution of column-based significance estimates for ClustalW SRP MSA's versus "MSA's" drawn from the null hypothesis.

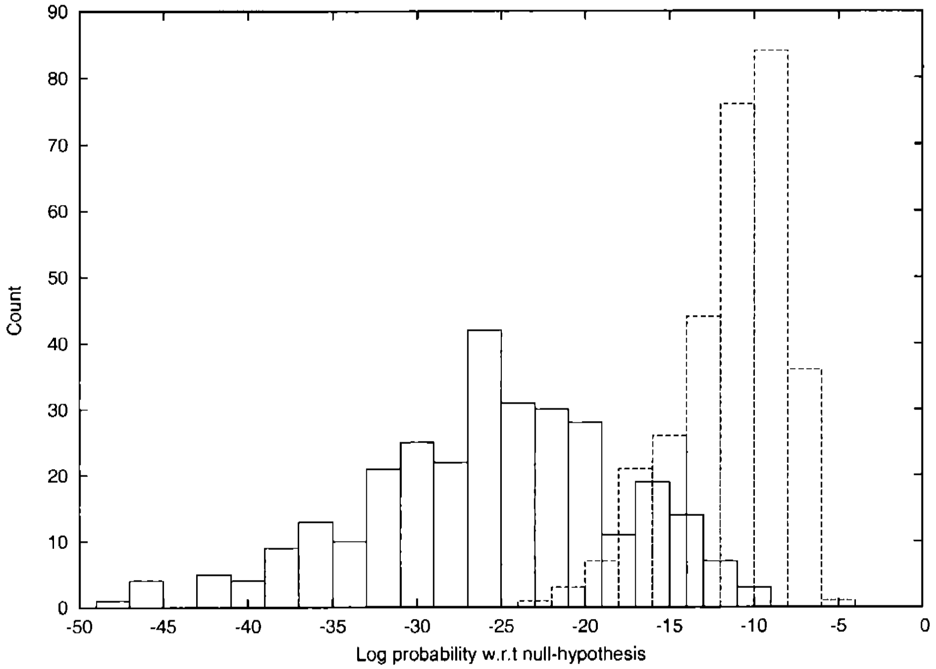

Figure 2-38: Distribution of column-based significance estimates for `ClustalW` SRP MSA's versus "MSA's" constructed by shuffling the columns of genuine MSA's.

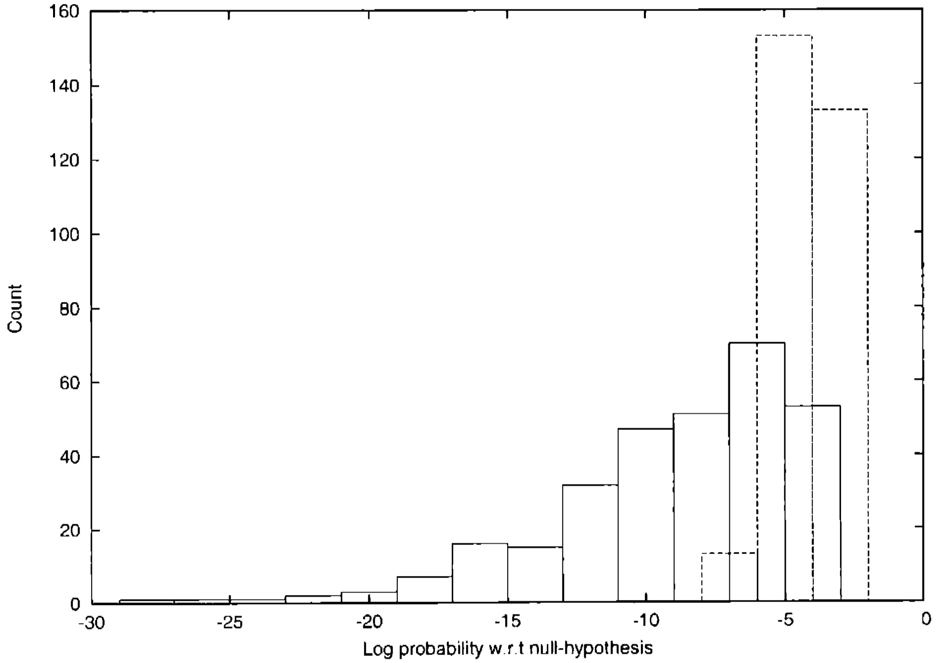

Figure 2-39: Distribution of column-based significance estimates for `ClustalW` SRP MSA's with strong local inter-row homology masked out versus "MSA's" drawn from the null hypothesis.

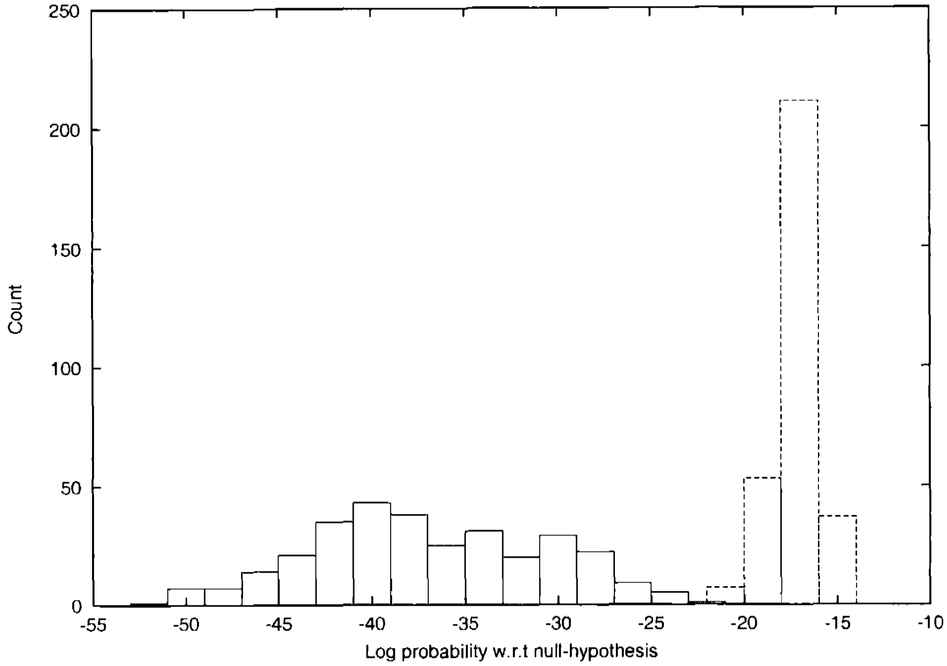

Figure 2-40: Distribution of tuple-based significance estimates for ClustalW RNaseP MSA's versus "MSA's" drawn from the null hypothesis.

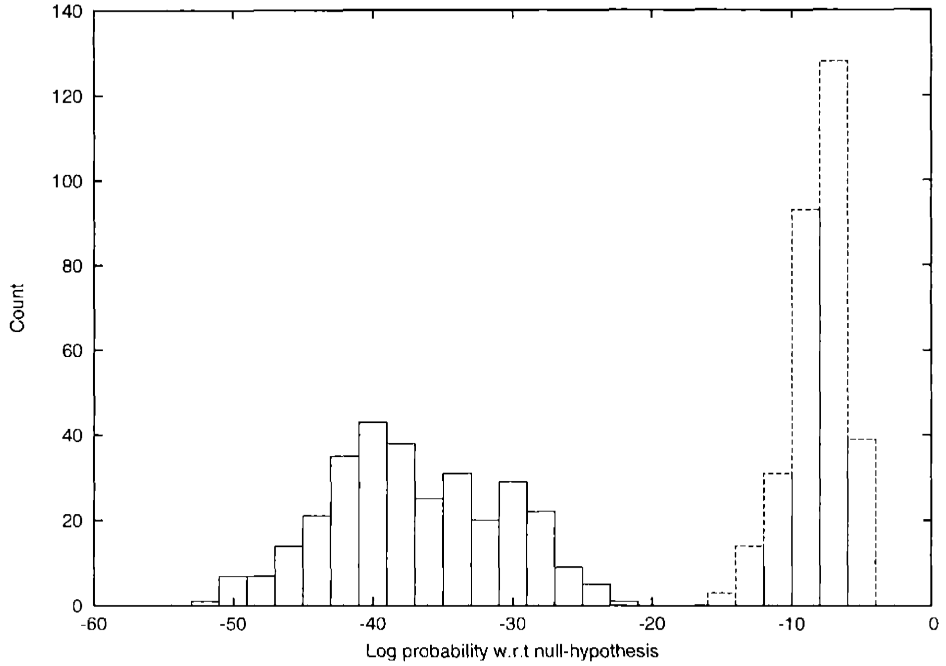

Figure 2-41: Distribution of tuple-based significance estimates for ClustalW RNaseP MSA's versus "MSA's" constructed by shuffling the columns of genuine MSA's.

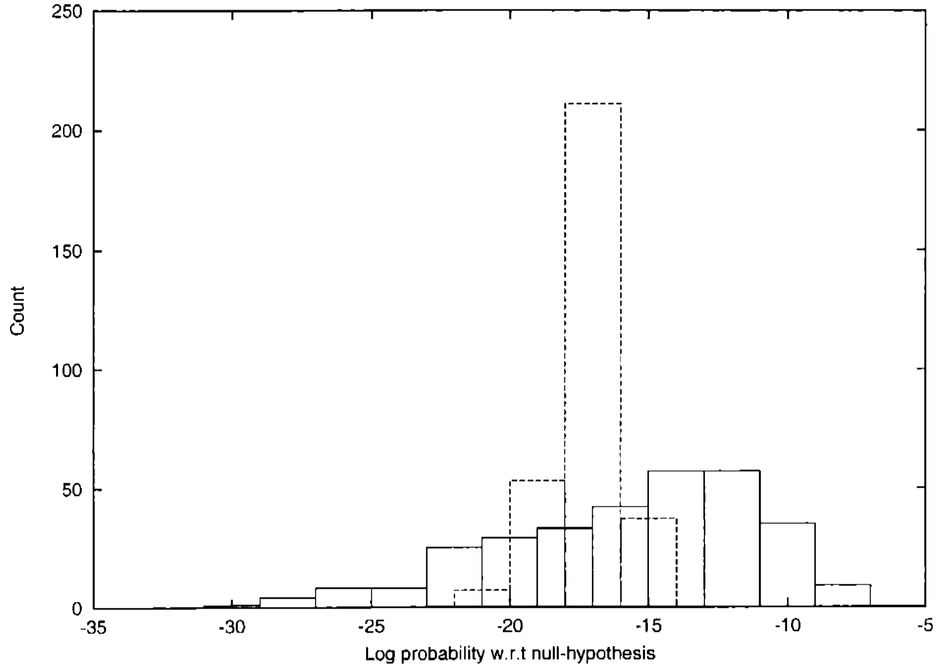

Figure 2-42: Distribution of tuple-based significance estimates for `ClustalW` RNaseP MSA's with strong local inter-row homology masked out versus "MSA's" drawn from the null hypothesis.

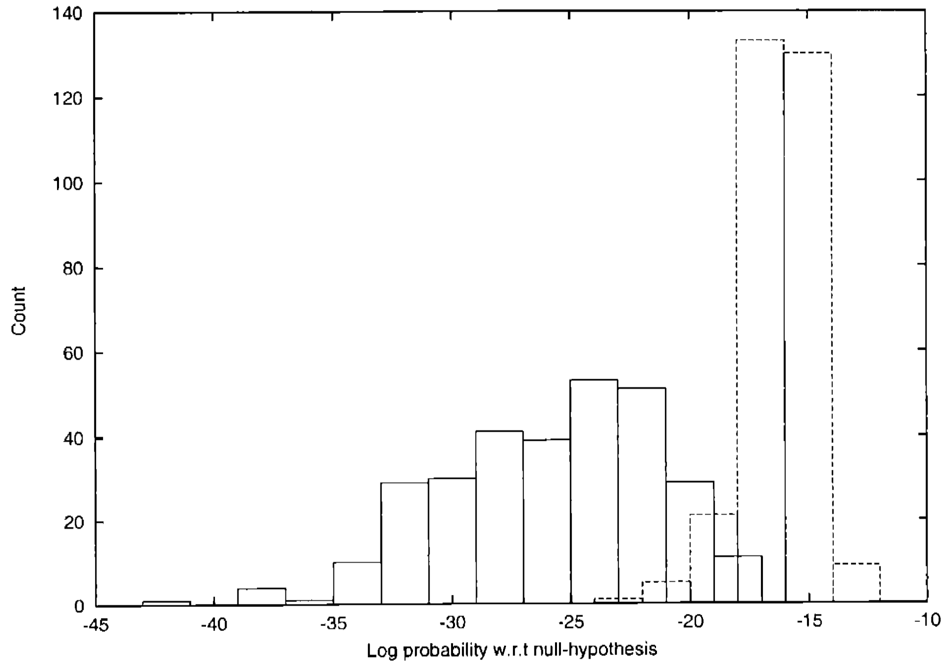

Figure 2-43: Distribution of tuple-based significance estimates for `ClustalW` SRP MSA's versus "MSA's" drawn from the null hypothesis.

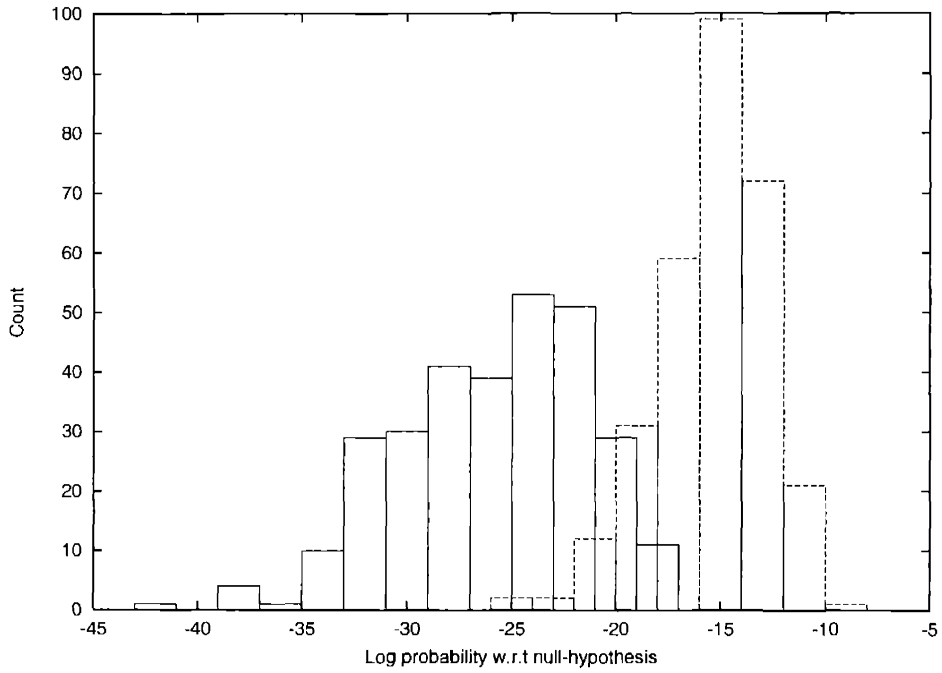

Figure 2-44: Distribution of tuple-based significance estimates for ClustalW SRP MSA's versus "MSA's" constructed by shuffling the columns of genuine MSA's.

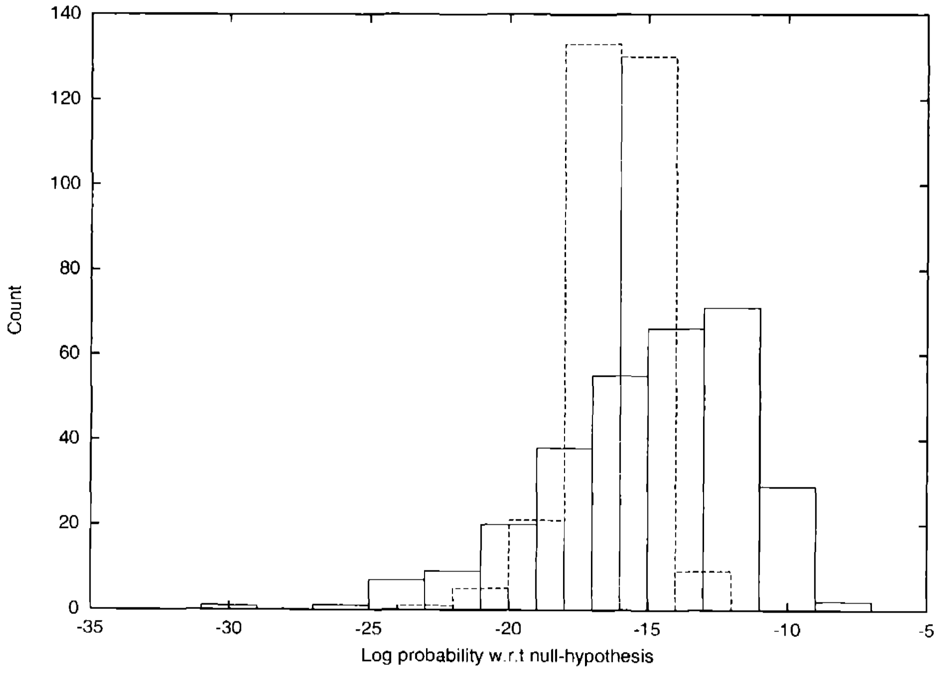

Figure 2-45: Distribution of tuple-based significance estimates for ClustalW SRP MSA's with strong local inter-row homology masked out versus "MSA's" drawn from the null hypothesis.

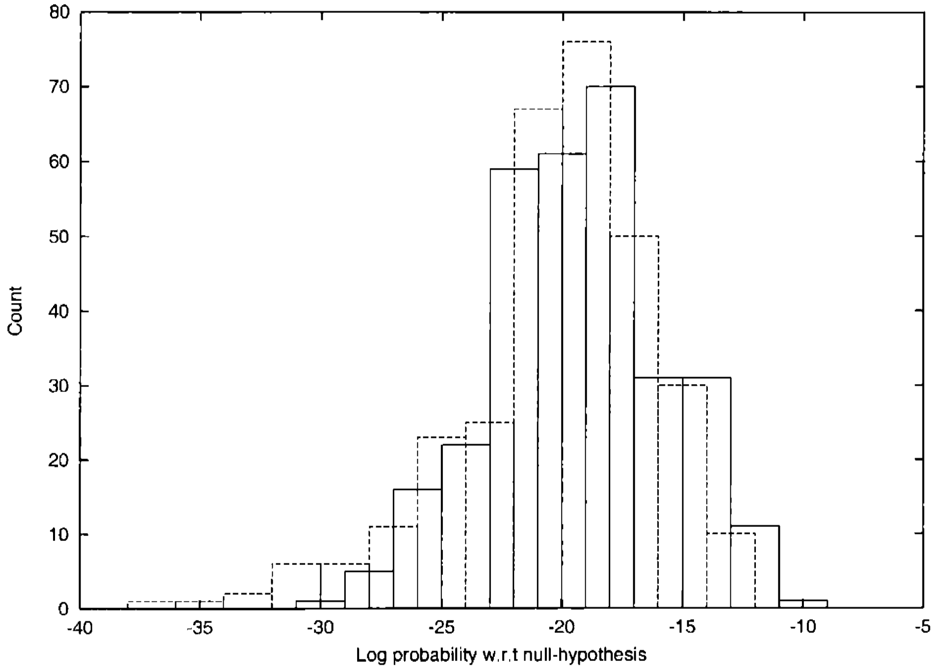

Figure 2-46: Distribution of helix-based significance estimates for `ClustalW` RNaseP MSA's versus "MSA's" drawn from the null hypothesis.

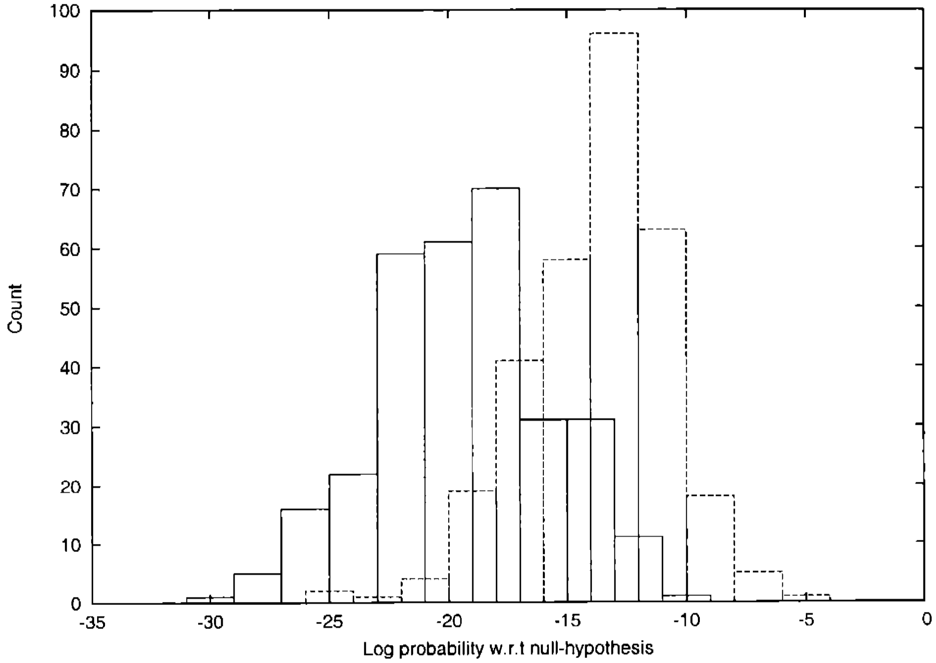

Figure 2-47: Distribution of helix-based significance estimates for `ClustalW` RNaseP MSA's versus "MSA's" constructed by shuffling the columns of genuine MSA's.

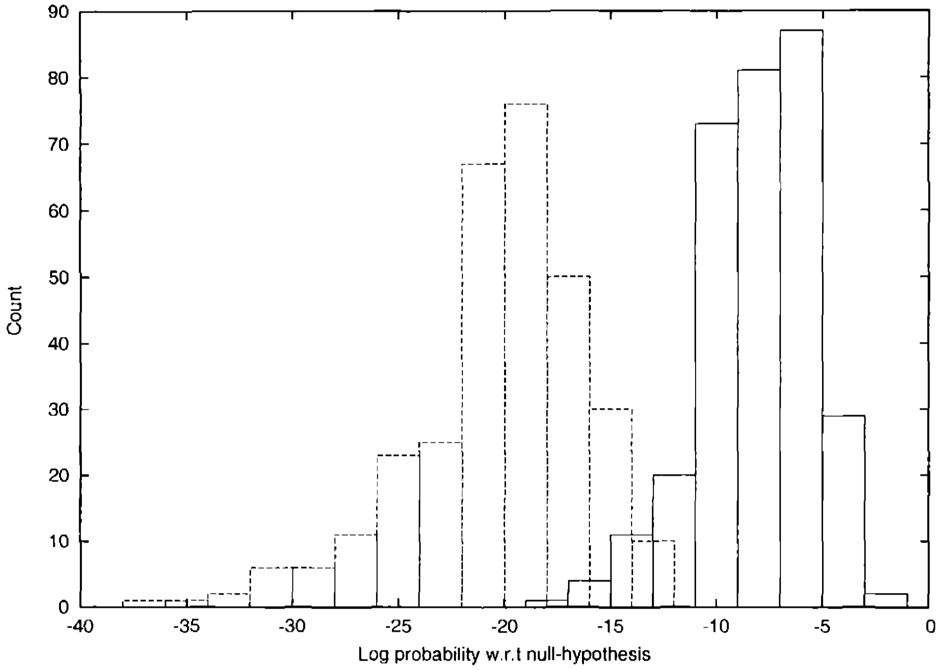

Figure 2-48: Distribution of helix-based significance estimates for ClustalW RNaseP MSA's with strong local inter-row homology masked out versus "MSA's" drawn from the null hypothesis.

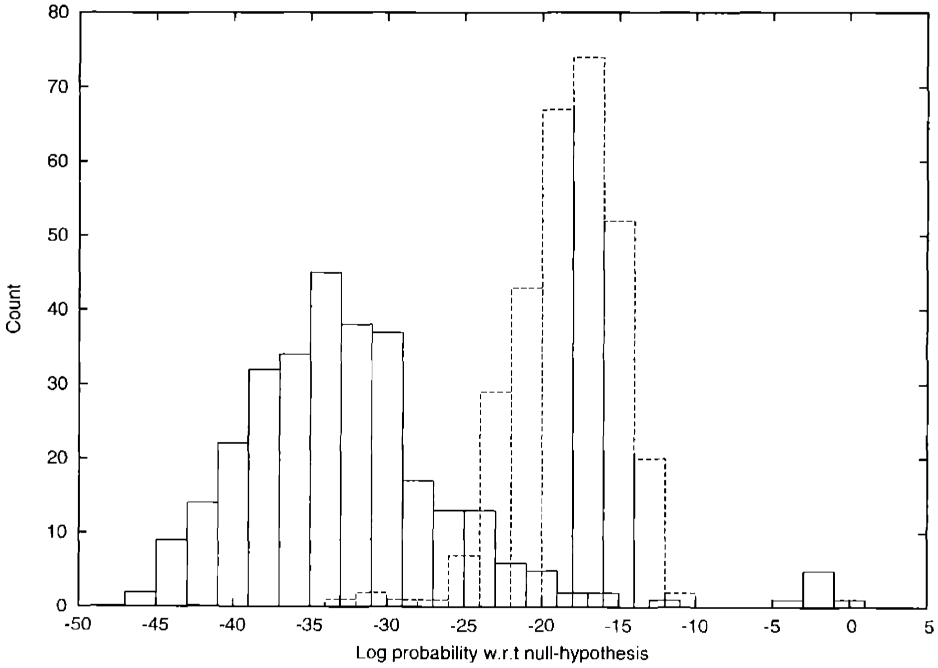

Figure 2-49: Distribution of helix-based significance estimates for ClustalW SRP MSA's versus "MSA's" drawn from the null hypothesis.

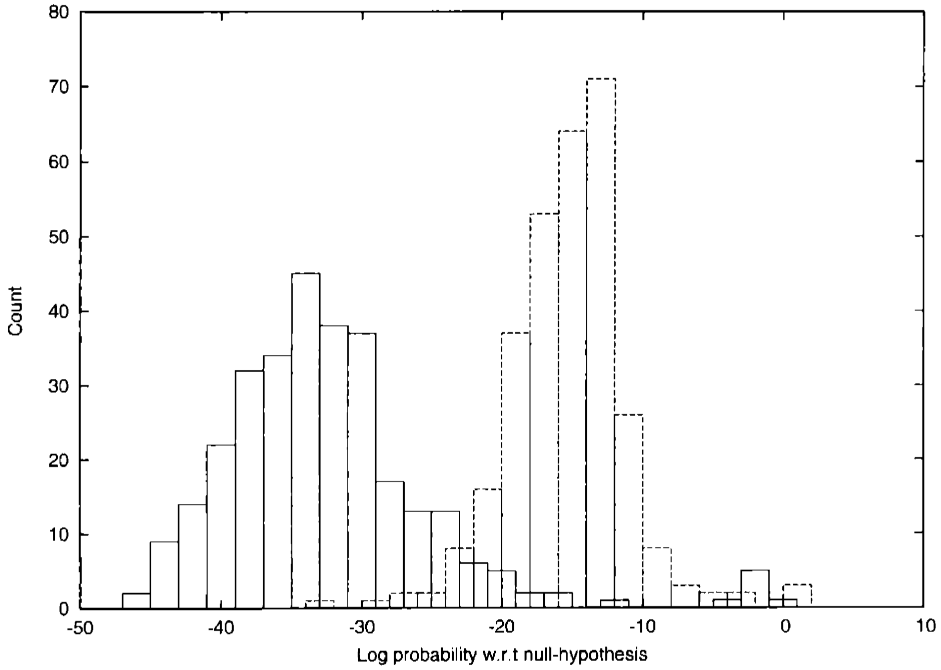

Figure 2-50: Distribution of helix-based significance estimates for ClustalW SRP MSA's versus "MSA's" constructed by shuffling the columns of genuine MSA's.

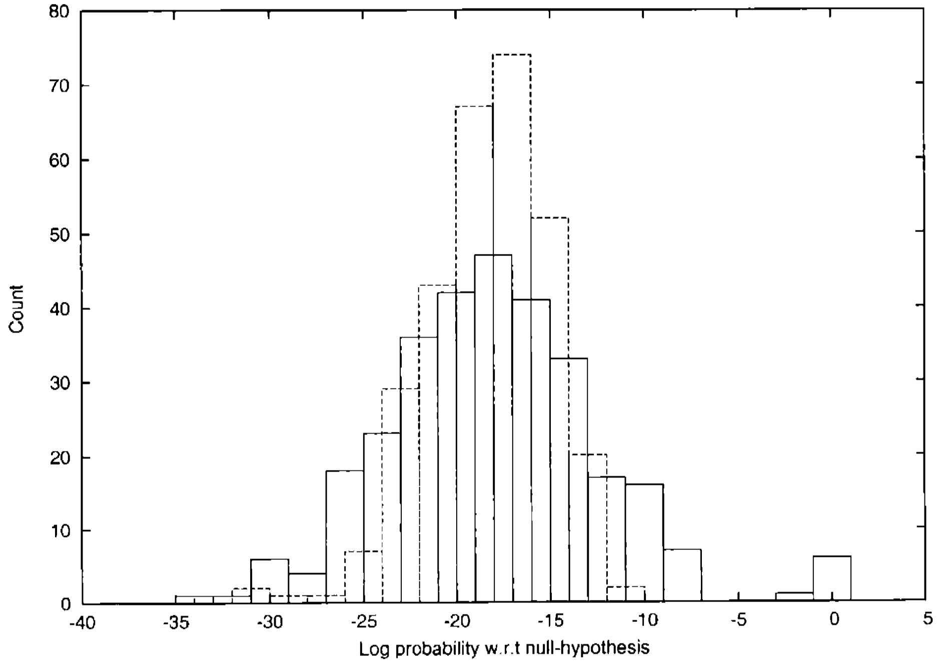

Figure 2-51: Distribution of helix-based significance estimates for ClustalW SRP MSA's with strong local inter-row homology masked out versus "MSA's" drawn from the null hypothesis.

2.2.5 Scan of bacterial genome for ncRNA's

Almost 100 bacterial genomes are available for download from `ftp://ftp.ncbi.nih.gov/genomes/Bacteria/` at the time of writing, and the wide range of phylogenetic distances between them make them excellent data for comparative methods. Following Rivas and Eddy's test of `QRNA` on *Salmonella Typhi* and *Escherichia Coli*, I have tested the techniques described above on MSA's of homologs generated from pairwise alignments of the *Buchnera* species' genome to the genomes of *Bacillus halodurans, Bacillus subtilis, Buchnera aphidicola schizaphis graminum, Escherichia Coli* strains K12, O157H7 and EDL933, *Helicobacter pylori* strains 26295 and J99, *Neisseria meningitidis* strains MC58 and Z2491, *Pasteurella multocida, Salmonella Typhi, Salmonella typhimurium* strain LT2, *Staphylococcus aureus* strains Mu50 and N315, and *Vibrio cholerae.*

The *Buchnera* genome was chosen as the central alignment subject simply because it is one of the smallest genomes on the list, and this speeds the alignment considerably. The entire list was chosen for reasonable phylogenetic closeness, following the phylogenetic information in [3].

Known coding regions in the *Buchnera sp.* genome were masked, and the resulting sequence was aligned to the other genomes in the list using `WU-BLAST` [24]. By ignoring coding regions the search missed mRNA's with secondary structure, but was faster by a factor of ten.

Alignments were collated by the position they matched in the *Buchnera sp.* genome. For each non-coding portion of the *Buchnera sp.* genome, the eight longest alignments to it were greedily chosen subject to the constraint that all pairwise alignments of the resulting nine sequences had nucleotide identities less than 85%. While the threshold of 85% was chosen fairly arbitrarily, it was necessary to ensure that the sequences were reasonably distinct from each other because very strong sequence similarity results in spurious correlations in the arrangements of complementary base pairs in the rows of the MSA.

If an MSA constructed in this way was longer than 300 base pairs, it was processed in portions of length 300 with an overlap of 150 base pairs. Thus an MSA whose sequences were 600 characters long was broken up into MSA's of length 300 starting at positions 0, 150, and 300

All 36 annotated RNA genes intersected an MSA with a significance greater than 20, and only two RNA genes failed to intersect one with significance greater than 25. There were 47 other non-overlapping regions containing an MSA with significance greater than 25. The boundaries of these regions are listed in table 2.2.5.

Some of the MSA's contained two known RNA's, and the significances of those were combined. Some of the significant pairs in MSA's intersecting known RNA genes did not lie within the annotated gene boundaries. Examples of typical significant MSA's intersecting known RNA's are given in table 2.2.5.

MSA start	MSA end	Prob.	RNA's	Pairs	Prob.'s
48417	48694	9.04e-16	(48488, 48560)	(48579, 48641)	6.84e-07
			(48576, 48647)	(48605, 48617)	1.43e-05
				(48591, 48599)	7.09e-05
				(48623, 48631)	1.30e-04
				(48426, 48571)	4.64e-06
				(48431, 48439)	2.43e-07
				(48491, 48554)	4.40e-06
				(48514, 48522)	6.90e-06
				(48496, 48509)	3.47e-06
				(48530, 48540)	3.64e-05
				(48444, 48459)	4.71e-05
				(48473, 48483)	1.99e-04
74459	74730	3.55e-32	(74471, 74543)	(74501, 74513)	3.31e-11
			(74572, 74644)	(74479, 74491)	2.19e-07
				(74567, 74721)	1.64e-06
				(74575, 74639)	6.11e-12
				(74589, 74626)	2.97e-07
				(74602, 74614)	1.38e-09
				(74648, 74683)	2.85e-06
				(74655, 74675)	1.27e-04
				(74662, 74670)	1.57e-04
				(74703, 74713)	1.56e-04
				(74518, 74556)	9.32e-06
				(74536, 74550)	1.48e-06
74595	74770	1.10e-11	(74572, 74644)	(74602, 74614)	1.38e-09
				(74726, 74740)	2.23e-05
				(74638, 74710)	6.46e-06
				(74648, 74683)	2.85e-06
				(74655, 74675)	1.27e-04
				(74662, 74670)	1.57e-04
				(74695, 74703)	2.26e-04

Table 2.5: Typical MSA's intersecting known ncRNA annotations in the *Buchnera* genome. "MSA start" and "MSA end" are the start and end indices of the portion of the *Buchnera* contig included in the MSA. "Prob." is the estimated probability of drawing an MSA exhibiting this degree of complementarity from the null hypothesis. "RNA's" is a list of the RNA gene annotations that the MSA intersects. "Pairs" and "Prob.'s" are respectively the start indices of the significant pairs of windows chosen from the MSA, and the estimated probabilities of drawing such pairs from a single sample of the null hypothesis.

(13954, 14329)	(20389, 20611)	(30501, 30764)	(35678, 36197)	(72186, 72409)
(101510, 101767)	(143341, 143859)	(179442, 179713)	(180129, 180540)	(181900, 182258)
(202652, 202829)	(211732, 212113)	(212128, 212317)	(223601, 223795)	(235400, 236082)
(251108, 251923)	(268257, 268464)	(276609, 277071)	(299704, 300048)	(327603, 328282)
(343173, 343424)	(347494, 347717)	(352802, 353014)	(355383, 355626)	(358251, 358466)
(358763, 358998)	(363409, 363632)	(370416, 370611)	(383497, 383763)	(384496, 384744)
(395201, 395459)	(411181, 411529)	(420568, 420778)	(425463, 425939)	(461608, 461849)
(526729, 526963)	(530995, 531190)	(538201, 538433)	(568591, 568850)	(579133, 579387)
(580393, 580794)	(581393, 582027)	(583462, 583689)	(609210, 609479)	(610238, 610471)
(611714, 612143)	(625408, 625682)			

Table 2.6: Predicted ncRNA gene regions

2.3 Future research

2.3.1 Further testing and refinement

- To check that the techniques described in section 2.2 are not fooled by correlations across homologs caused by something other than preservation of RNA secondary structure, their specificities need to be tested on genuine homologs that probably have no secondary structure. Known coding regions of mRNA's and repeat regions might be a good place to start.

- With clever indexing, it may be possible to identify columns exhibiting significant complementarity in linear time. If this turns out to be the case, it would be feasible to use the same technique to search a set of MSA's of known ncRNA orthologs for complementarity between positions in different genes, flagging probable *trans* base-pair interactions between the genes like those represented in figure 2-1.

2.3.2 Secondary structure prediction

Measuring complementarity by estimating its statistical significance appears to be a novel approach. The closest cognate in the literature that we are aware of is Hofacker's "covariance score" [27], which he uses in conjunction with free-energy minimization to predict consensus secondary structure of a set of ncRNA orthologs. It would be interesting to modify his secondary structure prediction algorithm to use significance estimates in place of his covariance score. Significance estimates might also be useful as a preprocessing step in secondary structure prediction, since its runtime grows as the square of the length of the sequence, while standard prediction algorithms must consider every trio of positions $i < j < k$ in the sequence, so their runtime grows cubically. If some base pairs can be confidently identified prior to applying standard algorithms, only triples that are consistent with those base pairs need to be considered, which can dramatically reduce the search space.

2.3.3 Searching for ncRNA genes with introns

Rivas and Eddy used QRNA to scan genomes of organisms that do not have introns. However, some known ncRNA genes such as XIST possess introns [6, 29]. No methods for identifying such genes is currently available. While QRNA has a "semi-Markov model" which it uses to predict the maximum-likelihood boundaries of ncRNA genes in the candidate orthologs it is passed, it makes no attempt to identify splice sites, and would run much too slowly even on a modest-sized candidate pair of orthologous genes with introns.

To compensate for this, one option is to use cDNA to obtain genes with concatenated exons, align those to another organism, concatenate the resulting alignments, and test the resulting alignments for complementary mutations using QRNA. In this way, the large set of cDNA's can be scanned, hopefully some novel ncRNA's can be identified.

I have implemented this pipeline, and it does find sequences that QRNA suggests are ncRNA genes. Because the set of known cDNA sequences is highly redundant, I searched the non-redundant set of cDNA's in the Unigene [40] database. Since long open reading frames are unlikely in non-coding genes, I further reduced the search by removing any entries with an open reading frame of length at least 90 codons. This left 50 965 entries which I BLAST'ed [24] against the Mouse genome [16]. In order to search for ncRNA genes with introns, I selected cDNA's that had multiple compatible alignments with expected frequencies of at most 0.01 separated by at most 10 000 base pairs. These alignments were concatenated, and those matches with 65% – 85% nucleotide identity were passed to QRNA, which reported 9 of the matches to be from alignments of ncRNA genes. Those 9 matches were BLAST'ed [2] against the NCBI nucleotide database [38] to check for strong homology to known genes. Five of the matches exhibited no such homology. Three of these matches were discarded as highly repetitious. The two remaining matches are in table 2.7.

In an attempt to expand this rather small collection, the entire set of Unigene representatives were blasted against the Mouse genome. This proved far more

```
Human cDNA #S1646953 against Mouse chromosome X contig MmX_30495_27

Mouse: 267694 ACCCATGGAATCGGGAAGGGACATGTCGCTGGACCGCTGCTGAGTCAGGGACTGATGCAT 267753
              | ||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Human:    252 AACCATGGAATCGGGCAGGGACATGTCGCTGGACCGCTGCTGAGTCAGGGACTGATGCAT 311

Mouse: 267754 GGTGAAAGCTGTCCAGTTGCCCCTGTCCCCGAAGCTGCTTCAAGAGCCCCTGGCTCAGGG 267813
              ||||||| ||| || |||||||| |||||||||||||||||||| ||||| ||||| |||
Human:    312 GGTGAAAACTGCCCGGTTGCCCCAGTCCCCGAAGCTGCTTCAAGCGCCCCCGGCTCTGGG 371

Mouse: 267814 CGCCCCGCAGAGCATCCTACTCTGCCGCTGCCT 267846
              |||||||||||||||||||||| || | |||||
Human:    372 CGCCCCGCAGAGCATCCTACTCCGCGGTTGCCT 404

Mouse: 266303 CCCTTGAACGAAATTTGTTGCGACTGAGTCTTTTGCAGGCAGCGGAGAGCTCACACTGGG 266362
              ||  |||||||  |||||||  |||||||||| ||| |||||  | ||| | | |  |||
Human:    111 CCGGTGAACGAGTTTTGTTGACACTGAGTCTTCTGCGGGCAGTAGGGAGATTATATAGGG 170

Mouse: 266363 C--TCAATCTCCT---TGCCCC-TCCCCT-GTGCATGTTCCAGCAAATTCATCGGAGAAG 266415
                 || |||| ||   |||||| ||| || | | ||||||| ||||||||||||||||||
Human:    171 GGTTCGATCTTCTCCCTGCCCCCTCCACTCGAG-ATGTTCCCGCAAATTCATCGGAGAAG 229

Mouse: 266416 AGCTGAGAAGACTGAATGGCAG 266437
              | ||||| ||||||| ||||||
Human:    230 ACCTGAGCAGACTGA-TGGCAG 250

Human cDNA #S1397703 against Mouse chromosome 11 contig Mm11_32669_27

Mouse: 44686 ATTCATATATCAGTTTCAAACAAGATGGGAATAAATCTATACATCACAAAAGGTACCTTG 44745
             || ||  || ||||||||||||| || ||   || | | ||| ||      | ||||||
Human:    11 ATACAG-TA-CAGTTTCAAACAAAATTGGGGCAACTTTGTACTTC------G-TACCTTA 61

Mouse: 44746 TTCAAGCTGGG-CTAGCACAGCCTGATTGACCTGTGAGGCTGACTTCAT 44793
             || || ||||| | | |||||| || |||   | |||   || ||||||
Human:    62 TTAAAACTGGGGCGAACACAGCGTGTTTGGT-T-TGATCTTGTCTTCAT 108

Mouse: 44896 ACAGAGTTACTTTCCA-T-CTGGGGAGTGATTCTCTATGAAGAATTGTAACTGTAAAACA 44953
             ||||| || |||||   | ||     | |||||| |  ||||||||||  ||||| |  |
Human:   267 ACAGA-TTCCTTTCAGGTACTTCTTTGGGATTCTGT--GAAGAATTGTGGCTGTACACTA 323

Mouse: 44954 AGATGTTTAATAGGAAG----CTCCTT---TTATGACATCATCAGACAGAGATTGTA 45003
              ||||||||||||||||    ||||||   ||||||||||| ||||||||||  |||
Human:   324 GGATGTTTAATAGGAAGGAAGCTCCTTCCTTTATGACATCACCAGACAGAGAG-GTA 379
```

Table 2.7: Non-contiguous matches of human cDNA to the Mouse genome flagged by QRNA as possible ncRNA orthologs.

fruitful, yielding 2268 raw BLAST matches and 55 matches after filtering by the same process described for human cDNA's. All but one of these matches was with a Rat cDNA transcript, a remarkable preponderance, given that only 75% of the BLAST matches involved Rat cDNA. Table 2.8 lists the candidate Unigene entries, and the corresponding portions of Mouse contigs.

QRNA also strongly flagged a number of matches that turned out to lie in known coding genes. These usually involved a match between a Rat cDNA and the $3'$ untranslated region of a coding gene. It would be interesting to systematically search orthologs of coding genes for the compensatory mutations associated with biologically significant RNA secondary structure, as there are a number of *cis*-acting mRNA's (see e.g. [23, 37, 31].) The genes QRNA flagged in this way were HLA-B in the Human Major Histocompatibility Complex, which codes for one of the surface proteins the immune system uses to recognize native cells [51], Human Rho GTPase activating protein 8, which is involved in the regulation of the actin cytoskeleton [12], and Munc13-4, which is believed be involved in regulating membrane traffic in the lungs [32]. Searches on PubMed [39] for these proteins revealed no articles postulating that the mRNA's of these molecules are *cis*-acting.

It is possible that this search would be improved by tuning QRNA's COD model to use mammalian codon mutation frequencies. The search was performed using the codon model QRNA ships with, which is presumably optimized for the bacterial genomes that Rivas and Eddy searched.

QRNA is currently too slow for large-scale searches for genes with introns

To search a target sequence for ncRNA genes whose cDNA has not been recorded and which may contain introns, some kind of exon assembly algorithm such as described in chapter 3 is needed. Scores for transcription start sites, splice sites and alignment to homologs can be used in the same fashion as when searching for coding genes. However, scores for the "exonness" of candidate exons need to be changed to reflect ncRNA secondary structure rather than protein structure. One straightforward approach would be to use a score based simply on the fidelity of alignments to

Unigene number	Mouse contig	Contig boundary indices
Rn#S363952	Mm1_25771_27	(160770, 160870), (161105, 161358)
Rn#S352137	MmX_32699_27	(214422, 214529), (214706, 214960)
Rn#S153994	Mm11_33611_27	(139013, 139345), (139622, 139695)
Rn#S8397	Mm11_32645_27	(89412, 89563), (89621, 89793)
Rn#S346768	Mm2_31545_27	(481370, 481183), (476719, 476486)
Rn#S331647	MmX_33906_27	(88745, 88548), (88242, 88103)
Rn#S364469	Mm11_31584_27	(52156, 52459), (52696, 52893)
Rn#S99922	MmX_26297_27	(241242, 241556), (241654, 241819)
Rn#S183908	Mm4_33748_27	(95565, 95636), (95787, 96123)
Rn#S202563	Mm3_33726_27	(100547, 100405), (100010, 99739)
Rn#S103326	Mm11_32645_27	(126006, 125846), (124408, 124171)
Rn#S361321	Mm11_31563_27	(386401, 386763), (386855, 387027)
Rn#S60948	Mm17_2761_27	(28911, 29007), (33081, 33242)
Rn#S149169	Mm10_27082_27	(113734, 113637), (113271, 113081)
Rn#S135958	Mm3_33573_27	(320525, 320759), (323496, 323599)
Rn#S270454	Mm4_33734_27	(139650, 139741), (141482, 141589)
Rn#S346381	Mm16_26295_27	(1754, 1939), (2381, 2427)
Rn#S329836	Mm11_33784_27	(115840, 115642), (115338, 115206), (110443, 110272)
Rn#S220512	Mm7_26253_27	(125004, 125088), (125501, 125724)
Rn#S186253	MmX_33916_27	(97654, 97783), (97891, 98150)
Rn#S189469	Mm2_33706_27	(6874, 6560), (6471, 6398)
Rn#S213733	Mm3_33722_27	(69869, 70039), (70040, 70137)
Rn#S342964	Mm11_33613_27	(170925, 170668), (170649, 170600)
Rn#S363472	Mm13_33622_27	(173793, 173922), (179239, 179442)
Rn#S275483	Mm11_33781_27	(158911, 159098), (159249, 159477)
Rn#S121414	Mm1_32609_27	(261673, 261835), (262011, 262106)
Rn#S167450	Mm17_30483_27	(17137, 17220), (17440, 17569)
Rn#S181440	Mm11_33842_27	(221725, 221498), (219543, 219261)
Rn#S331811	Mm7_2748_27	(6911, 6827), (6681, 6534)
Rn#S219465	Mm11_31561_27	(174778, 174571), (174389, 174301)
Rn#S231960	Mm11_33842_27	(6767, 7010), (7043, 7100)
Rn#S244366	Mm11_31586_27	(322046, 322224), (329017, 329344)
Rn#S247280	Mm2_31545_27	(395455, 395361), (395279, 395192)
Rn#S93586	Mm11_33784_27	(188329, 188125), (188070, 187912)
Rn#S266706	Mm6_32337_27	(5225, 5411), (11597, 11799)
Rn#S299413	Mm2_31545_27	(443961, 443632), (443364, 443241)
Rn#S322323	Mm13_31593_27	(61520, 61279), (61201, 61065), (61025, 60919)
Rn#S325922	Mm11_33829_27	(160679, 160757), (160787, 160957)
Rn#S334941	Mm2_33711_27	(75862, 76121), (76332, 76521)
Os#S15730	Mm11_33776_27	(128215, 128274), (134434, 134498)
Rn#S341712	Mm11_31584_27	(420856, 420709), (419778, 419636)
Rn#S361828	Mm11_33617_27	(144077, 144374), (152979, 153097)
Rn#S363456	Mm3_26367_27	(89205, 89295), (91576, 91753)
Rn#S360597	Mm4_33742_27	(97271, 97024), (96112, 95882)
Rn#S354848	Mm17_30483_27	(204662, 204905), (204932, 205192)
Rn#S323624	Mm11_33862_27	(112119, 112494), (112847, 113050)
Rn#S305824	Mm11_26339_27	(183942, 184178), (184365, 184654)
Rn#S358387	Mm1_25695_27	(60259, 59933), (58662, 58480)
Rn#S358387	Mm1_25680_27	(198459, 198787), (200056, 200240)
Rn#S80041	Mm6_32337_27	(37537, 37230), (28699, 28647)
Rn#S55695	Mm4_2900_27	(42388, 42184), (42135, 42100)
Rn#S179344	Mm11_33775_27	(161115, 160950), (160948, 160530)
Rn#S54091	Mm16_26295_27	(137040, 136804), (136728, 136629)
Rn#S176913	Mm11_33848_27	(47802, 47630), (44677, 44493)
Rn#S345794	Mm11_31572_27	(18183, 17811), (16601, 16561)
Rn#S253488	Mm2_33718_27	(169200, 169160), (169132, 168926)
Rn#S348943	Mm13_30491_27	(165454, 165873), (170718, 170873)

Table 2.8: Non-contiguous matches between Unigene cDNA's and the Mouse genome flagged by QRNA as ncRNA orthologs. In cases where the first contig boundary exceeds the second, the match was to the contig portion's reverse complement.

Contig name	Exon boundaries in parse
Mm11_31586	107, 1013, 1676, 2611, 3399, 3469, 5859, 6182, 7327, 7489, 8845, 10160, 10446, 10906, 12010, 12338, 13232, 15602, 16995, 17799, 20432, 22658, 23553, 23879, 25549, 26150, 26377, 26439
Mm11_31587	5258, 5800, 8648, 8955, 9154, 10609
Mm11_32645	996, 1038, 1633, 3453, 3941, 4618, 4730, 4942, 5629, 6739, 6938, 7119, 7281, 8239, 10110, 10435, 10493, 11869

Table 2.9: Plausible gene parses for the Mouse contigs table 2.8 were generated using the exon assembly algorithm described in chapter 3 and the splice-site scores in chapter 4. Candidate exon interiors were scored for homology to Rat contigs chosen for strong matches to the corresponding Rat cDNA's in table 2.8. The homology score used was the number of identical nucleotides in alignments to the corresponding rat contig divided by 100.

candidate orthologs. This approach is vulnerable to spurious cross-genome matches, and quite error-prone. That could be compensated for by computing a large number of high-scoring parses and choosing from among these the one whose concatenated exons and alignments have the highest score according to QRNA. Another would be to concatenate strongly aligned regions between the two orthologs with consistent orientations and positions, and run QRNA on the concatenation, hoping that it can pick the alignment out of that. It turns out that both of these approaches fail because the alignments obtained by either method are simply too long for QRNA to process at a reasonable speed. Rat contigs [52] with strong homology to the Rat cDNA's in table 2.8 form alignments to the corresponding Mouse contigs that are thousands of nucleotides long. Table 2.9 shows some parses gene parses generated from those mouse contigs.

Chapter 3

Assembling global gene predictions

Existing tests for local features, including those in this thesis, are quite inaccurate. One can compensate for this to a certain extent by choosing from the predicted local features subsets which are consistent with the overall structure common to all genes and which have high total probabilities. This means finding a transcription start site, followed by an alternating sequence of 5′ and 3′ splice sites:

$$\underbrace{\texttt{ATG}\ldots\texttt{GT}}_{\text{exon}}\ldots\underbrace{\texttt{AG}\ldots\texttt{GT}}_{\text{exon}}\ldots\underbrace{\texttt{AG}\ldots}_{\text{exon}}$$

In the case of genes which code for protein, the concatenation of the exons also has to be free of stop codons. In other words, if the concatenation is broken up into groups of three, starting with the initial ATG, there should be no instances of the triples TAA, TAG, or TGA.

Here is a rough description of the algorithm for finding the best gene prediction from a set of local gene features (for more information, see, e.g. [8]): one scans across the predicted local features in linear order, keeping a list of partial predictions that have been formed so far. A new parse is added each time a translation start site (ATG) is encountered, and the score of the ATG is recorded as the score of this parse. Each time a 3′ splice site (AG) is encountered, a new parse is generated for each one in the list that ends with a 5′ splice site (GT). The scores of the new parses are the scores of their prefix parses plus the score of the 3′ splice site, plus possibly a score for the "intronness" of the intervening intron. The highest-scoring of these newly generated parses is added to the list of parses. Each time a GT is encountered, new parses are generated from those in the list that the GT can consistently be added to. The highest-scoring of these is is added to the list, where again the score of a new parse is the score of the prefix parse, plus the score of the GT, plus a score for the "exonness" of the resulting exon. When all local features have been processed in this way, the highest-scoring parse in the list is the highest-scoring parse given those local features and scores.

When searching for coding genes, the constraint that the concatenation of exons should contain no stop codons is very helpful. It increases the accuracy of gene predictions, because it reduces the number of candidate parses to pick the correct parse from. Taking it into account also dramatically improves the efficiency of the algorithm. Without it, its time to run scales roughly bilinearly in the the number of candidate GT's and the total number of AG's and ATG's, as every GT has to be compared to a parse ending in every prior AG and ATG. However, there is almost always a stop-codon just a few hundred base pairs prior to any given GT that precludes appending it to a a partial parse that ended much earlier, and by keeping track of this it is possible for the search to scale linearly on average in the total number of predicted gene features.

For ncRNA genes, this constraint does not apply, so an alternative algorithm will have to be used. To deal with this case, the following algorithm would be appropriate. It is a divide-and-conquer algorithm that takes advantage of the fact that the scoring function is local and additive, so given a subsequence of DNA, it is possible to determine the optimal parse within that subsequence knowing only the state of the global optimal parse at the boundaries of the subsequence.

Let $S = n_1 \dots n_N$ be a sequence of nucleotides of length N. Because our scores for local features are local and additive, optimality of a parse on S is a local property, in the sense that its restriction to a subsequence $S' = n_j n_{j+l} \dots n_k$ is optimal among subparses that satisfy the boundary conditions imposed by the state of the parse at the edges of S'. Thus we can compute the optimal parse on the entirety of S by breaking it up into constant-length portions, computing optimal subparses on each portion for each set of possible boundary conditions, and gluing these optimal subparses together. Here is a more formal description. We begin with some definitions. Let $\mathcal{T}, \mathcal{A}, \mathcal{G}, \mathcal{E} \subset \{1, \ldots, N\}$ be disjoint sets of positions of candidate transcription start sites, $3'$ splice sites, $5'$ splice sites and transcription end sites, respectively. A *parse* of S is a sequence $1 \leq f_1 \leq \cdots \leq f_l \leq N$ with $f_1 \in \mathcal{T}$, $f_2 \in \mathcal{G}$, $f_l \in \mathcal{E}$, $f_{l-1} \in \mathcal{A}$, $f_i \in \mathcal{A}$ if $f_{i-1} \in \mathcal{G}$, and $f_i \in \mathcal{G}$ if $f_{i-1} \in \mathcal{A}$. The *score* of such a parse is

$$TSS(f_1) + TSE(f_l) + \Sigma_{2,4,6,\ldots<l} G(f_i) + \Sigma_{1,3,5,\ldots<l} A(f_i) + \Sigma_{1,3,5,\ldots l} E(f_i, f_{i+1}).$$

Here, $TSS(f_1)$ is the score for f_1 as a transcription start site, $TSE(f_l)$ the score for f_l as a transcription end site, $G(f_i)$ the score for f_i as a $5'$ splice site, $A(f_i)$ the score for f_i as a $3'$ splice site, and $E(f_i, f_{i+1})$ the score for the region between f_i and f_{i+1} inclusive as an exon.

For a subsequence $S' = n_I \ldots n_{I+N'-1}$, a *subparse* is a sequence $I \leq f_1' < \ldots < f_{l'}' \leq I + N' - 1$ with $f_1' \in \mathcal{T} \cup \mathcal{A} \cup \mathcal{G}$, $f_{l'}' \in \mathcal{A} \cup \mathcal{G} \cup \mathcal{E}$, $f_2', \ldots, f_{l'-1}' \in \mathcal{A} \cup \mathcal{G}$, $f_i' \in \mathcal{G}$ if $f_{i-1}' \in \mathcal{A} \cup \mathcal{T}$, $f_i' \in \mathcal{A}$ if $f_{i-1}' \in \mathcal{G}$, and $f_{l'-1}' \in \mathcal{A}$ if $f_{l'}' \in \mathcal{E}$. The score of such a subparse is the sum of the scores of the f_i's, plus the sum of the scores of any internal exons, plus $E(I, f_1)$ if $f_1 \in \mathcal{G}$, and $E(f_{l'}', I + N' - 1)$ if $f_{l'} \in \mathcal{A} \cup \mathcal{T}$. For $\mathcal{F}_1, \mathcal{F}_2 \in \{\mathcal{A}, \mathcal{G}, \mathcal{T}, \mathcal{E}\}$, let $P(S', \mathcal{F}_1, \mathcal{F}_2)$ denote a highest-score subparse $f_1', \ldots, f_{l'}'$ in S' with $f_1' \in \mathcal{F}_1, f_{l'}' \in \mathcal{F}_2$. Such a parse can be computed in $O((N')^2)$ time using a simple modification to the standard dynamic programming algorithm for exon assembly. Let $P(S', \emptyset, \emptyset)$ denote an empty parse, with score 0.

Suppose U and V are adjacent subsequences. Denote their concatenation by UV. If P, Q are subparses of U, V respectively, and they can sensibly be concatenated, denote this concatenated subparse of UV by PQ.

Here is the algorithm for exon assembly. Assume $N > 5000$. Divide S into adjacent subsequences of length 1000, with the last sequence possibly shorter than that: $S_1 = n_1 \ldots n_{1000}, S_2 = n_{1001} \ldots n_{2000}, \ldots$. Assume for simplicity that this results in an even number of subsequences. Add an empty subsequence to the list to ensure this, if need be. The algorithm recursively operates on the set of subsequences as follows:

(1) For each S_i, compute the following parses: $P(S_i, \mathcal{T}, \mathcal{E})$, $P(S_i, \mathcal{A}, \mathcal{E})$, $P(S_i, \mathcal{T}, \mathcal{A})$,

$P(S_i, \mathcal{A}, \mathcal{A})$, $P(S_i, \mathcal{T}, \mathcal{G})$, $P(S_i, \mathcal{A}, \mathcal{G})$.

(2) Next, for i odd, let $\overline{S_i} = S_i S_{i+1}$ denote their concatenation. Because all the scores are local, the optimal subparses for this subsequence listed in step 1 can be constructed from those for S_i and S_{i+1}. For instance $P(\overline{S_i}, \mathcal{T}, \mathcal{E})$ is the highest scoring of the following concatenated parses: $P(S_i, \mathcal{T}, \mathcal{E})P(S_{i+1}, \emptyset, \emptyset)$, $P(S_i, \emptyset, \emptyset)P(S_{i+1}, \mathcal{T}, \mathcal{E})$, $P(S_i, \mathcal{T}, \mathcal{T})P(S_{i+1}, \mathcal{A}, \mathcal{E})$, $P(S_i, \mathcal{T}, \mathcal{A})P(S_{i+1}, \mathcal{G}, \mathcal{E})$, $P(S_i, \mathcal{T}, \mathcal{A})P(S_{i+1}, \mathcal{E}, \mathcal{E})$, $P(S_i, \mathcal{T}, \mathcal{G})P(S_{i+1}, \mathcal{A}, \mathcal{E})$.

(3) If there is more than one subsequence in $\{\overline{S_i}\}$, replace the S_i's with the $\overline{S_i}$'s, ensure that there are an even number of subsequences in the list by adding an empty one to the end of the list if need be, and repeat from step 2. If $S_1 = S$, the highest scoring parse is $P(S_1, \mathcal{T}, \mathcal{E})$

Note that $O(N \log N)$ such concatenations are required to fully assemble S, and step 1 takes time $O(N)$. Hence the run time of the algorithm grows as $O(N \log N)$.

The algorithm can easily be modified to keep track of the highest-scoring n parses by simply keeping the n highest-scoring parses at each stage. The time complexity of the algorithm grows linearly with n.

Chapter 4

Statistical tests for local gene features

This chapter describes new techniques for identifying transcription start sites and splice sites. The proteins that trigger transcription and splicing exhibit affinities to the DNA/RNA sequences comprising these features, but as mentioned in the Introduction, such affinities cannot be determined *a priori* yet.

Identifying the characteristic patterns of gene features is a frustrating problem—presumably, a gene's entire behavior is specified by its DNA sequence, so whatever these patterns are, they are staring back at us from the thousands of known genes, completely accessible if only we were clever enough to see them. However, the patterns might be very complex, at least in some cases. One phenomenon that suggests this notion is the strong conservation of splice sites and transcription start sites in some relatively distant orthologs. If the crucial properties of those sites are too delicate to admit significant changes over time from point mutations to the underlying sequence, then they may be too complex to capture using the simple statistical models that

4.1 Tests based on non-contiguous patterns

The tests for local features in GENSCAN [8, 9] and FGENE [50] are both based on statistical models with a relatively small number of parameters. At the time they were devised, this was necessary, because the set of known genes that could be used for training was quite small. However, a far larger set of genes has since been recorded, and correspondingly richer tests can be trained from them. To take advantage of this, I devised a class of tests that capture correlations between non-adjacent nucleotide positions. The resulting tests have no clear probabilistic model behind them, but seem to be effective in capturing previously unused patterns in gene features.

The main problem with a test that captures arbitrary non-adjacent correlations is that the statistically significant patterns that result can overlap in arbitrary ways, and a useful probabilistic model for estimating the probability of overlapping patterns has proven elusive. The somewhat arbitrary approach taken here is to greedily choose the most statistically significant pattern in the sequence, then the next most that doesn't overlap that pattern, then the next most, and so on. These events are then treated as independent. For instance, suppose the following sequence is to be tested as a potential 5′ splice site:

	0123456789
0-9	AAGCCCTGAG
10-11	GAGGTGCCAT
20-29	CTCCCTCTTT
30-35	TCCAG

First all of the patterns it contains are listed and sorted by their statistical significance in the training sequences (just how this statistical significance is calculated will be described later.)

Pattern	Start index	Score
CCT*T*T*TT	22	4.53
T*C*T*TT*T	21	4.50
C*T*TTTT	23	4.50
C*TCTTT	23	4.40
T*C*T*T*TT	21	4.21
TCTTTTC	25	4.18
300 patterns elided...		
C*T**T*T	23	2.33
T**C*C*T**C	21	2.32
TC*CC*C*T*T	19	2.32

First the CCT*T*T*TT pattern is chosen. The next most significant pattern, T*C*T*TT*T, is discarded, because it overlaps pattern already chosen, the patterns are examined in order of decreasing scores until another consistent pattern is found; T**C*C*T**C. This process is repeated until the list of patterns is exhausted or the entire sequence to be tested is covered:

Pattern	Start index	Score
CCT*T*T*TT	22	4.53
T**C*C*T**C	21	2.32
A**TG**AT	11	2.19
C*CTGA	3	1.43
C*{14}C	17	1.04
C***C	16	0.46
C*{7}G	4	-0.35
A*{8}G	1	-0.46
A*{12}G	0	-0.64

(Here $*\{n\}$ is used to denote n wild cards in succession.)

Then the procedure is repeated, using the *lowest* scoring sequences:

Pattern	Start index	Score
GG*G**A	12	-3.18
G**C*G**G	2	-1.25
G*A	9	-0.55
C*{11}C	4	0.18
T*{10}C	6	0.20
A*{13}T	0	0.24
A*{11}C	8	0.26
C*{18}C	3	0.34
CC	23	0.38
A*{17}T	1	0.52
CT	26	0.64
CC	31	0.66
T******T	21	1.18
TT	29	1.48

Finally, to obtain the overall score for the splice site, the scores for the patterns from these two lists are summed.

4.1.1 Calculating the statistical significance of the individual patterns

Patterns' frequencies in intergenic DNA were used as a null hypothesis. The per-event probability of them appearing was estimated from these frequencies. If the probability of a pattern's observed frequency in the training data was less than 10^{-7} with respect to the null hypothesis, the pattern was considered sufficiently significant for this testing scheme. The score used for sorting patterns was

$$\log\left(\frac{T_p}{T}\right) - \log\left(\frac{N_p}{N}\right)$$

where T_p is the number of times the given pattern was seen in the training data, T is the number of times a pattern with the nucleotide offsets seen in the given pattern was seen in the training data, and N_p, N are the corresponding values for the null hypothesis data.

The way the "nucleotide offsets" were counted varied depending on the types of patterns expected. For splice sites, where it was expected that the features significant to gene expression would have fixed nucleotide positions relative to some easily identifiable anchor (AG or GT,) T was the number of times a pattern (with *any* set of nucleotides) was seen at the positions occupied by the pattern under consideration. For the null hypothesis counts in these cases, N was the number of times a pattern with that set of nucleotide offsets (again, with any set of nucleotides occupying those relative positions.) For transcription start sites, where there is no obvious anchor in candidate sites, patterns with a given nucleotide offset were counted equivalently no matter where they appeared in the training sequences. However, the accuracy of the resulting test for transcription start sites has since been exceeded by other research [13].

4.2 Performance of splice site tests

For both $3'$ and $5'$ sites, training sets were taken from the gene parses in the GBPRI [5]. Null hypothesis sets were taken from random intergenic data.

4.2.1 The $3'$ test

The test was trained on the fifty nucleotides preceding the AG signaling resumption of coding sequence. The patterns lengths and maximal gap lengths allowed were respectively two nucleotides with gaps of length less than 20, three nucleotides with gaps of length at most 5, four nucleotides with gaps of length at most 4, five nucleotides with gaps of length at most 3, and six nucleotides with gaps of length at most 2.

Two hundred sequences each were removed at random from the training and null sets prior to training. A cut-off score was chosen based on the performance of the test on some of the data it was trained on. These were the results for the data excluded from training:

True positives	191
False negatives	9
True negatives	179
False positives	21
Sensitivity	95.5%
Specificity	90.1%

The $3'$ splice-site test described in [8, 9] was tested on 600104 nucleotides of data, about one-sixteenth of which would have provided false AG's to be used as testing data. At a sensitivity of 95%, it found 5, 397 false positives. At the rate demonstrated above, this new test would find about 3, 700 false positives.

4.2.2 The $5'$ test

The test used the same set of patterns as the $3'$ test. It was trained on the three nucleotides preceding and the fifty nucleotides following the GT signaling the end of coding. Two hundred sequences each were removed from the training and null sets prior to training. The results from those sequences were:

True positives	190
False negatives	10
True negatives	192
False positives	8
Sensitivity	95.0%
Specificity	96.0%

The corresponding MDD-based test in [8, 9] had 3382 false positives, so both tests have about the same false positive rate. However, they use different data, and are thus somewhat independent, as the scatter diagram in figure 4-1 shows. Thus it is possible to combine them into slightly a more accurate test.

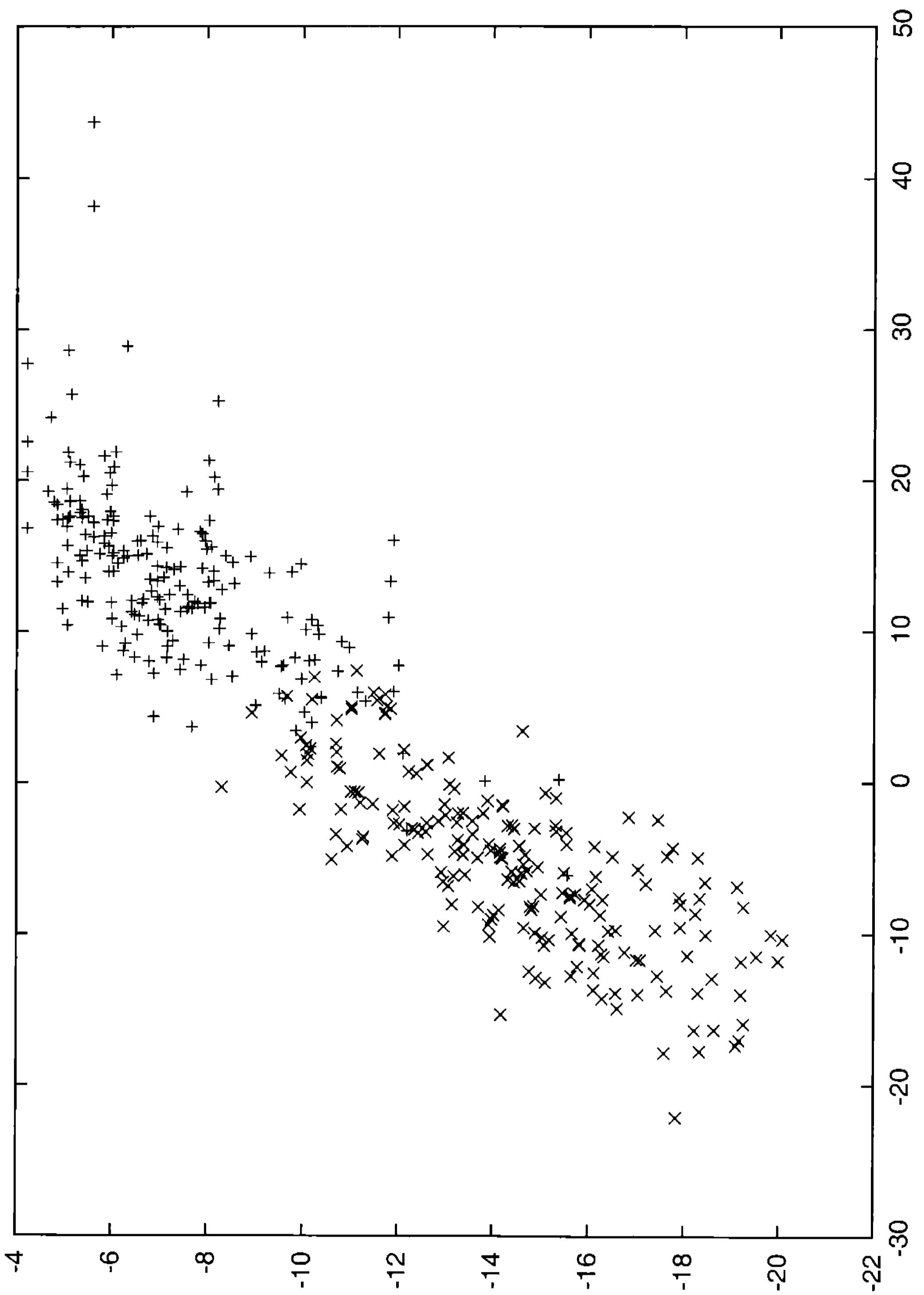

Figure 4-1: Scatter plot of pattern-based scores against GENSCAN's [8, 9] MDD-based scores. Plus signs mark scores of actual 5′ sites, crosses mark scores of randomly chosen sequences. The X-axis is the pattern-based score, the Y-axis is the MDD-based score.

Chapter 5

Cartesian Currents

5.1 Preliminaries

This section contains some notations, conventions and basic results which will be used in this chapter.

For details on standard notations and results from Geometric Measure Theory see for instance [48] or [17].

5.1.1 Notation for Euclidean spaces

Fix two natural numbers n, N. Throughout this chapter, let Ω be a bounded open subset of $\mathbb{R}^n$ with smooth boundary. Let $\pi : \mathbb{R}^n \times \mathbb{R}^N \to \mathbb{R}^n$, $\hat{\pi} : \mathbb{R}^n \times \mathbb{R}^N \to \mathbb{R}^N$ be the projections onto the first and second factors, respectively. Let $e_1, \dots, e_n$ denote the standard basis of $\mathbb{R}^n$.

Suppose $m \in \mathbb{N}$, $p \in \mathbb{R}^m$ and $\epsilon > 0$. Then as usual, let

$$B_\epsilon(p) = \{x \in \mathbb{R}^m \mid |x - p| < \epsilon\}.$$

5.1.2 Forms and currents

This section outlines the basic definitions of currents.

An open subset of $\mathbb{R}^m$ will have the standard orientation unless there is an explicit indication to the contrary.

A $\mathcal{H}^k$-measurable set $M \subset \mathbb{R}^m$ is called *countably k-rectifiable* if it is $\mathcal{H}^k$-almost all contained in a countable union of k-dimensional C^1-submanifolds of $\mathbb{R}^m$.

Given an open set $G \subset \mathbb{R}^m$, the space of smooth differential k-forms which are compactly supported in G has a standard locally convex topology (see [48, §26] for the definition of this topology). Denote by $\mathcal{D}_k(G)$ the dual of this space; its members are called k-currents.

Composing currents with the exterior derivative operator induces a map $\partial : \mathcal{D}_k(G) \to \mathcal{D}_{k-1}(G)$. Given a current $V \in \mathcal{D}_k(G)$, the value of ∂V on a differential $(k-1)$-form ω is given by

$$\partial V(\omega) = V(d\omega)$$

where $d\omega$ is the exterior derivative of ω. The map ∂ is called the *boundary map.* We will refer to ∂V as the *boundary* of V. If $\partial V = 0$ then we say that V is *boundaryless.*

Suppose $M \subset \mathbb{R}^m$ is a countably k-rectifiable set with finite $\mathcal{H}^k$-measure. Then for $\mathcal{H}^k$-almost every $x \in M$, there is an approximate k-dimensional tangent plane to M denoted by T_xM ([48, 11.4-6]). Suppose $\theta : M \to \mathbb{Z}$ is a $\mathcal{H}^k$-measurable function with locally finite L^1-norm. Suppose $\xi : M \to \bigwedge_k(\mathbb{R}^m)$ is a $\mathcal{H}^k$-measurable map such that for $\mathcal{H}^k$-almost every $x \in M$, $\xi(x) \in \bigwedge_k(T_xM) \subset \bigwedge_k(\mathbb{R}^m)$ and $\xi(x)$ is a *unit vector* (with respect to the inner product on $\bigwedge_k(\mathbb{R}^m)$ induced by the standard inner product on $\mathbb{R}^m$.) Then $\tau(M, \theta, \xi)$ denotes the current whose value on a differential k-form ω having compact support is given by

$$\tau(M, \theta, \xi)(\omega) = \int_M \langle \omega, \xi \rangle \theta d\mathcal{H}^k.$$

Such a current is called *integer multiplicity rectifiable.*

The *mass* of such a current is given by

$$I\!M(\tau(M, \theta, \xi)) = \int_M |\theta| d\mathcal{H}^k.$$

Definition 5.1.1. The set of all rectifiable k-currents in $\mathcal{D}_k(G)$ with finite mass will be denoted by $\mathcal{R}^k(G)$

Given $T = \tau(M, \theta, \xi) \in \mathcal{R}^k(G)$ and a $\mathcal{H}^k$-measurable set $S \subset \mathbb{R}^m$, denote by $T \lfloor S$ the current $\tau(M \cap S, \theta, \xi)$.

For open sets $U \subset \mathbb{R}^P$, $V \subset \mathbb{R}^Q$, and a smooth proper map $f : U \to V$, the pullback by f of a form $\omega \in \mathcal{D}^n(V)$ is the form $f^{\#}(\omega) = (Df)^{\#}(w \circ f) \in \mathcal{D}^n(U)$. For a current $S \in \mathcal{D}^n(U)$ its *push forward* by a proper function f is denoted by $f_\#(S)$, and its value on a form $\omega \in \mathcal{D}^n(V)$ is $S(f^\#\omega)$. By [48, 26.25], the push-forward may also be defined when f is merely Lipschitz and proper. For a k-current T,

$$I\!M(f_\#(T)) \leq \text{ess sup}|Df|^k I\!M(T).$$

The graph of a function

Suppose $u : \Omega \to \mathbb{R}^N$. Then the *graph* of u is the set

$$G_u = \{(x, u(x)) \mid x \in \Omega\}$$

Suppose further that u represents a member of $BV(\Omega)$ and its graph has locally finite $\mathcal{H}^n$-measure (in particular, this is so if $u \in C^1(\Omega, \mathbb{R}^N)$). Then G_u is a countably n-rectifiable set (see [22, Proposition 1]). Suppose that wherever the tangent plane of G_u is defined it is not orthogonal to $\Omega \times \{0\}$. Give it the orientation which makes orthogonal projection onto Ω orientation preserving. Let

$$\overrightarrow{G_u} : G_u \to \bigwedge\nolimits_n (\mathbb{R}^n \times \mathbb{R}^N)$$

be a map corresponding to this orientation. In other words, if $x \in G_u$ and $T_x G_u$ exists then $\overrightarrow{G_u}(x)$ is the unique unit vector in $\bigwedge_n(T_x G_u) \subset \bigwedge_n(\mathbb{R}^n \times \mathbb{R}^N)$ for which $\pi_\# \overrightarrow{G_u}(x) = \alpha e_1 \wedge \ldots \wedge e_n$, with $\alpha > 0$.

Differential n-forms can be integrated over G_u, inducing

$$[\![G_u]\!] = \tau(G_u, \theta, \overrightarrow{G_u}) \in \mathcal{R}^n(\Omega \times \mathbb{R}^N).$$

5.1.3 Cartesian currents

Definition 5.1.2.

$$\begin{aligned} \mathsf{cart}^* \ (\Omega, \mathbb{R}^N) \ &= \ \{T = \tau(M, \theta, \xi) \mid \pi_\# T = [\![\Omega]\!] \text{ and for } \mathcal{H}^n\text{-almost all} \\ &\qquad z \in M, \ \pi_\#\xi(z) = \alpha e_1 \wedge \ldots \wedge e_n \text{ with } \alpha \geq 0\} \end{aligned}$$

Suppose $T = \tau(M, \theta, \xi) \in \mathsf{cart}^* \ (\Omega, \mathbb{R}^N)$. Without loss of generality, assume that for all $z \in M$, $T_z M$ exists, $\xi(z) \in \wedge^n T_z M$ is a unit vector, and $\frac{\pi_\#\xi(z)}{e_1 \wedge \ldots \wedge e_n} \geq 0$. The co-area formula (see [21, Theorem 2, Section 3] and [20, Theorem 5, Section 2]) implies that M contains the graph of a function of bounded variation. In particular, for almost every $x \in \Omega$ there is a unique point $(x, y_x) \in M$, and so a function $u_T : \Omega \to \mathbb{R}^N$ can be defined almost everywhere in Ω by

$$u_T(x) = y_x.$$

and then u_T can be extended to all of Ω by defining it to be zero elsewhere. This function's graph G_{u_T} is $\mathcal{H}^n$-almost all contained in M, and moreover $\pi(M \setminus G_{u_T})$ is $\mathcal{H}^n$-null. It is also easy to see that $\theta|_{\{z \in G_{u_T} \mid T_z M \not\perp \Omega \times \{0\}\}} = 1$ $\mathcal{H}^n$-almost everywhere.

Conversely, suppose $T = \tau(M, \theta, \xi) \in \mathcal{R}^n$ is boundaryless and that there exists a function $u_T : \Omega \to \mathbb{R}^N$ representing an element of $BV(\Omega, \mathbb{R}^N)$ such that $\pi(M \Delta G_u)$ is $\mathcal{H}^n$-null and such that $\theta|_{G_u} = 1$ $\mathcal{H}^n$-almost everywhere. If $\pi_{\#} T = [\![\Omega]\!]$, then it is not hard to check that $T \in \mathsf{cart}^*\ (\Omega, \mathbb{R}^N)$. In particular, if $u \in C^\infty(\Omega, \mathbb{R}^N)$ and $\mathcal{H}^n(G_u) < \infty$ then $[\![G_u]\!] \in \mathsf{cart}^*\ (\Omega, \mathbb{R}^N)$.

Redefining a function on a set of measure zero may result in a function whose graph is not $\mathcal{H}^n$-measurable, or which differs from the graph of the original function by a set which is not $\mathcal{H}^n$-null. For instance, suppose that $N \geq n$. Take an uncountable set $U \subset \Omega$ of measure zero and construct a bijection $b : U \to \mathbb{R}^N$. Then given any function $u : \Omega \to \mathbb{R}^N$, redefine it to equal b on U. This results in a function whose graph has orthogonal projection onto all of $\mathbb{R}^N$. Thus the graph of this new function cannot have finite $\mathcal{H}^n$-measure.

The following five definitions are based on those given in [21].

Definition 5.1.3. The function u_T discussed immediately above is called the *underlying function* of T.

Definition 5.1.4. Suppose $T \in \mathsf{cart}^*\ (\Omega, \mathbb{R}^N)$ has underlying function u_T. Then the C-norm of T is

$$\|T\|_{\mathsf{C}} = I\!\!M(T) + \|u_T\|_{L^1}$$

Definition 5.1.5. $\mathsf{cart}(\Omega, \mathbb{R}^N) \subset \mathsf{cart}^*\ (\Omega, \mathbb{R}^N)$ is the set of $T \in \mathsf{cart}^*\ (\Omega, \mathbb{R}^N)$ for which $\|T\|_{\mathsf{C}} < \infty$.

Definition 5.1.6. Suppose $(T_k)_{k=1}^{\infty} \subset \mathsf{cart}(\Omega, \mathbb{R}^N)$. Then $T \in \mathsf{cart}(\Omega, \mathbb{R}^N)$ is called the *weak* C-*limit* of the sequence (T_k) if

(1) $\|T_k\|_{\mathsf{C}}$ is bounded independently of k,

(2) $T_k \rightharpoonup T$ (i.e. for each differential n-form ω with compact support in $\Omega \times \mathbb{R}^N$, $T_k(\omega) \to T(\omega)$).

This convergence is denoted by

$$T_k \overset{\mathsf{C}}{\rightharpoonup} T.$$

The space $\mathsf{cart}(\Omega, \mathbb{R}^N)$ is closed under this notion of convergence [22].

Definition 5.1.7. The space $\mathsf{Cart}(\Omega, \mathbb{R}^N)$ is the smallest subset of $\mathsf{cart}(\Omega, \mathbb{R}^N)$ which is closed under sequential weak C-convergence (see Definition 5.1.6) and which contains

$$\{[G_u] \mid u \in C^1(\Omega, \mathbb{R}^N) \text{ and } \|[G_u]\|_{\mathsf{C}} < \infty\}.$$

In other words, $\mathsf{Cart}(\Omega, \mathbb{R}^N)$ is the closure under weak C-convergence of the set of currents given by integration over graphs of smooth functions.

The following Lemmas will be useful in the sequel.

Lemma 5.1.8. *Suppose $T \in \mathsf{cart}(\Omega, \mathbb{R}^N)$, x_i the ith co-ordinate function. Then for almost all $x \in \mathbb{R}$ and any connected component $C \subset \Omega \cap \{x_i = x\}$,*

$$\langle T, x_i, x \rangle \lfloor C \times \mathbb{R}^N \in \mathsf{cart}(C, \mathbb{R}^N).$$

Proof. Obvious.

□

Lemma 5.1.9. *Let $v : \Omega \to \mathbb{R}^N$ be a function for which*

(1) There is an open $V \subset\subset \Omega$ and a compact $K \subset V$ such that $v|_V$ is Lipschitz and $v|_{\Omega \setminus K}$ is smooth.

(2) $\mathcal{H}^n(G_v)$ is finite.

Then $[G_v] \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$.

Proof. Choose $\phi \in C^\infty(\Omega)$ for which spt $\phi \subset V$ and $\phi \equiv 1$ in some neighborhood of K. Choose a sequence $(w_l) \subset C^\infty(\Omega, \mathbb{R}^N)$ with a uniform Lipschitz bound in V and which converges uniformly to v in V. Define

$$v_l(x) = (1 - \phi(x))v(x) + \phi(x)w_l(x)$$

Then v_l is smooth, and its restriction to V satisfies a Lipschitz bound B which is independent of l. Thus

$$\begin{aligned} \mathcal{H}^n(G_{v_l}) &\le \mathcal{H}^n(G_u \cap ((\Omega \setminus V) \times \mathbb{R}^N)) + \mathcal{H}^n(G_{v_l} \cap V) \\ &\le \mathcal{H}^n(G_u \cap ((\Omega \setminus V) \times \mathbb{R}^N)) + C(1 + B^n)\mathcal{H}^n(V) \end{aligned}$$

and clearly $v_l \to v$ uniformly. Thus $[G_{v_l}] \overset{\mathsf{C}}{\rightharpoonup} [G_v]$.

□

Lemma 5.1.10. *Suppose $\mathcal{W}$ is the smallest set closed under weak* C*-convergence of the set $\{[G_u] \mid u \in V\}$ for some $V \subset C^\infty(\Omega, \mathbb{R}^N)$. Suppose $F : \mathcal{W} \to \mathsf{cart}(\Omega, \mathbb{R}^N)$ is a map which is continuous with respect to the mass norm, and that there is some $c, d > 0$ such that for $T \in \mathcal{W}$, $\|F(T)\|_{\mathsf{C}} \leq c\|T\|_{\mathsf{C}} + d$. Finally, suppose that for $u \in V$, $F([G_u]) \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$. Then $F(\mathcal{W}) \subset \mathsf{Cart}(\Omega, \mathbb{R}^N)$.*

Proof. This is an easy transfinite induction.

□

5.2 Bubbling of vertical currents

In this section, it is shown that if $V \in \mathcal{R}^n(\mathbb{R}^N)$, $x \in \mathbb{R}^N$ and $T \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$, then $T + [\{x\}] \times V \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$.

Lemma 5.2.1. *Let $f : S^n \to \mathbb{R}^N$ be a Lipschitz map. Let $u \in C^\infty(\Omega, \mathbb{R}^N)$ be such that $[G_u]$ has finite mass. Then for $x \in \Omega$, $[G_u] + [\{x\}] \times f_\#[S^n] \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$.*

Proof. Let $w : \overline{B_1(x)} \to S^n$ be a smooth surjection constant on $\partial B_1(x)$ and one-to-one elsewhere which is orientation preserving with the standard orientation on S^n. Let

$$\tilde{f}(x') = \begin{cases} f \circ w(x') & \text{if } |x' - x| < 1 \\ (|x' - x| - 1)u(x) + (2 - |x' - x|)f \circ w(\frac{x'}{|x'|}) & \text{if } |x' - x| \in [1, 2) \\ u(x) & \text{if } |x' - x| \geq 2 \end{cases}$$

Let $\rho : \mathbb{R}^n \to \mathbb{R}$ be a smooth function such that $\rho|_{B_2(0)} = 1$, and $\rho|_{\mathbb{R}^n \setminus B_3(0)} = 0$. Let

$$u_k(x') = \begin{cases} u((1 - \rho(k|x - x'|))x' + \rho(k|x - x'|)x) & \text{if } x' \notin B_{\frac{2}{k}}(x) \\ \tilde{f}(kx') & \text{if } x' \in B_{\frac{2}{k}}(x). \end{cases}$$

Then by lemma 5.1.9, for sufficiently large k, $[G_{u_k}] \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$. Since $[G_{u_k}] \xrightarrow{\mathsf{C}} [G_u] + f_\#(S^n)$, the result follows.

□

Corollary 5.2.2. *Let f be as in the previous lemma, $T \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$. Then $T + [\{x\}] \times f_\#([S^n]) \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$.*

Proof. This follows from the above Lemma and Lemma 5.1.10, with $F(T) = T + [\{x\}] \times f_\#([S^n]) \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$, and $\mathcal{W} = \mathsf{Cart}(\Omega, \mathbb{R}^N)$.

□

Corollary 5.2.3. *Suppose $V \in \mathcal{R}^n(\mathbb{R}^N)$ a boundaryless current, $x \in \Omega$, $T \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$. Then $T + [\{x\}] \times V \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$.*

Proof. By the Weak Polyhedral Approximation Theorem [48, 30.2], for each l there is a sequence of finite sums $W_h = \Sigma_i f_{h,i\#}([S^n])$ converging weakly to V and having uniformly bounded masses controlled by the mass of V. On the other hand, by the above corollary, $T + \llbracket\{x\}\rrbracket \times W_h \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$, and since $T + \llbracket\{x\}\rrbracket \times W_h \overset{C}{\rightharpoonup} T + V$, it follows that $T + \llbracket\{x\}\rrbracket \times V \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$.

□

5.3 The proof of the conjecture for one-dimensional domains

In this section, Ω will be an open interval of the real line. Mollifications of the underlying function u_T for any $T \in \mathsf{cart}(\Omega, \mathbb{R}^N)$ give a sequence of smooth functions $u_l \overset{BV}{\rightharpoonup} u_T$. Since Ω is 1-dimensional, the lengths of their graphs are controlled by their BV-norms, and hence are uniformly bounded. Thus the graphs give a sequence of cartesian currents with uniformly bounded masses, so by taking a subsequence, it may be assumed that $(\llbracket G_{u_l}\rrbracket)$ converges weakly to some $S \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$ for which $u_S = u_T$ almost everywhere. Write $T = \tau(M_T, \theta_T, \xi_T)$, $S = \tau(M_S, \theta_S, \xi_S)$. By the definition of a rectifiable current as an integral, any current $C = \tau(M, \theta, \xi) \in \mathcal{R}^n(U)$ is unchanged if M is replaced by the subset of all $x \in M$ for which $\theta(x) \neq 0$, the approximate tangent plane $T_x M$ exists, and $\xi(x)$ is a unit vector in $\wedge^n T_x M$. Assume this replacement has been made for M_S and M_T. Then there is $G \subset M_T \cap M_S$ such that for almost all $x \in \Omega$, $\{x\} \times \mathbb{R}^N \cap (M_T \cup M_S) = \{(x, u_T(x))\} = \{(x, u_S(x))\}$ and $\xi_T(x, u_T(x)) = \xi_S(x, u_S(x)) = 1$. Hence, $T - S$ can be written as an integral over $(M_T \cup M_S) \setminus G$ and $\mathcal{H}^1(\pi(M_T \cup M_S) \setminus G) = 0$. Thus it suffices to show that any current $\tau(M, \theta, \xi)$ such that $\mathcal{H}^1(\pi(M)) = 0$, can be weakly approximated by currents of the form considered in Corollary 5.2.3:

Lemma 5.3.1. *Let $V = \tau(M, \theta, \xi) \in \mathcal{R}^n(\Omega \times \mathbb{R}^N)$ be a boundaryless current with $\mathcal{H}^1(\pi(M)) = 0$. Then V is the weak limit of currents of the form*

$$\Sigma_i \llbracket\{q_i\}\rrbracket \times V_i$$

where $q_i \in \Omega$, and $V_i \in \mathcal{R}^n(\mathbb{R}^N)$

Proof. For $\rho > 0$, let $\{x_1 < \ldots < x_m\} \subset \Omega$ be a set of points such that $\Omega \subset \bigcup B_\rho(x_i)$, and for $i = 1 \ldots m$,

$$\langle V, x, x_i \rangle = 0, \text{ and } \partial(V \lfloor \{x \in \Omega \mid x < x_i)\}) = 0$$

(see [48, 28.4,28.5].) Let $p\colon\Omega\setminus\{x_1,\dots x_m\}\to\{x_1,\dots x_m\}$ be the map

$$p(x)=\begin{cases}\min\{x_i\mid x_i>x\} & \text{if } x<x_m\\ x_m & \text{otherwise}\end{cases}$$

Then define

$$p_\#(V)=\lim_{j\to\infty}p_\#(V\lfloor\Omega\setminus\bigcup_{i=1}^{m}(x_i-d_j,x_i+d_j)),$$

where $d_j\to 0$ is a sequence such that for all j, $\partial(V\lfloor\{x\in\Omega\mid x<d_j)\})=0$. If $h(t,x,y)=((1-t)x+tp(x),y)$, then $h_\#(V)$ can be defined similarly, and $\partial h_\#(V)=V-p_\#(V)$. Since $I\!M(h_\#(V))\le\rho I\!M(V)$, and ρ was arbitrary, a sequence of currents of the required form converging to V can be constructed this way.

□

5.4 Proof of conjecture for higher dimensions

The remainder of the proof is by induction on the dimension of Ω.

By Lemma 5.5.1 the closure under C-weak convergence of the set

$$\mathcal{F}=\{S\in\mathsf{cart}(\Omega,\mathbb{R}^N)\mid\text{spt }(S-[\![G_0]\!])\subset\subset\Omega\times\mathbb{R}^N\}$$

is all of $\mathsf{cart}(\Omega,\mathbb{R}^N)$. Thus it suffices to prove the conjecture for $T\in\mathcal{F}$. Fix such T. Then there is open $\Omega_T\subset\subset\Omega$ and $R>0$ such that T may be written as $T=\tau(M,\theta,\xi)$, with $M\subset\Omega_T\times B_R(0)$. Without loss of generality, it may also be assumed that for any $z\in M$, the weak tangent plane T_zM exists, $\theta(z)\ne 0$, and $\xi(z)\in\wedge^nT_zM$ and is a unit vector. Also, fix a null-set $N_T\subset\Omega$ such that for $x\in\Omega\setminus N_T$, the set $M\cap(\{x\}\times\mathbb{R}^N)$ contains just one point y_x and $\pi_\#(\xi(x,y_x))=\alpha e_1\wedge\ldots\wedge e_n$, with $\alpha\ge 0$.

Lemma 5.4.1. *Suppose $T\in\mathcal{F}$. For C an open n-cube, let $\mathcal{P}_C$ be the set of faces of C. Then for sufficiently small ρ, there is a collection of n-cubes*

$$\mathcal{T}=\{A_l=\alpha+\rho z_l+(0,\rho)^n\mid z_l\in\mathbb{Z},A_l\subset\Omega,\ \mathcal{H}^{n-1}(N_T\cap\partial A_l)=0,\\ \text{and }\partial(T\lfloor A_l\times\mathbb{R}^N)\in\mathcal{R}^{n-1}(\Omega\times\mathbb{R}^N)\}_l$$

such that

$$\Sigma_l\rho(I\!M(\partial(T\lfloor(A_l\times\mathbb{R}^N))+\Sigma_{F\in\mathcal{P}_{A_l}}\|u_{T_F}\|_{L^1})\le 2n\|T\|_{\mathsf{C}},$$

where if $F\in\mathcal{P}_{A_l}$, then $T_F=\partial(T\lfloor A_l\times\mathbb{R}^N)\lfloor F\times\mathbb{R}^N$. That this is an element of $\mathsf{cart}(F,\mathbb{R}^N)$ *follows from the requirement $\mathcal{H}^{n-1}(N_T\cap\partial A_l)=0$.*

Proof. Choose ρ small enough that if

$$J = \left\{\rho z + (0,\rho)^n \mid z \in \mathbb{Z}^n, \rho z + (0,\rho z)^n \subset \{x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) > \sqrt{n}\rho\}\right\},$$

then $\Omega_T \subset \bigcup_{A \in J} \overline{A}$. For $x \in (0,\rho)^n$, let $J_x = \{x + A \mid A \in J\}$.

Let G be the full-measure set of all x such that the result of Lemma 5.4.1 holds. There is $\alpha \in G$ such that

$$\rho(\Sigma_{A \in J_\alpha} I\!M(\partial(T \lfloor A)) + \Sigma_{F \in \mathcal{P}_\alpha} \|u_{T_F}|_F\|_{L^1}) \leq 2n\|T\|_{\mathsf{C}}.$$

To find such α, let x_i denote the ith co-ordinate function and let

$$\begin{aligned} B \;=\; \bigcap_{i \in \{1,\ldots n\}, j \in \mathbb{Z}} & \{\, r \in (0,\rho) \mid \langle T, x_i, r + j\rho\rangle = \partial(T\lfloor\{x_i \geq r + j\rho\}) \text{ and} \\ & \langle T, x_i, r + j\rho\rangle \in \mathsf{cart}(\{x_i = r + j\rho\}, \mathbb{R}^N)\,\}\,. \end{aligned}$$

Note that B has full measure [48, 28.1]. Then

$$\begin{aligned} \|T\|_{\mathsf{C}} \;&\geq\; \int_{-\infty}^{\infty} \|\langle T, x_i, r\rangle\|_{\mathsf{C}}\, dr \\ &=\; \Sigma_{j \in \mathbb{Z}} \int_B \|\langle T, x_i, r + j\rho\rangle\|_{\mathsf{C}}\, dr \\ &=\; \int_B \Sigma_{j \in \mathbb{Z}} \|\langle T, x_i, r + j\rho\rangle\|_{\mathsf{C}}\, dr, \end{aligned}$$

and so summing this inequality over i, the set of $(x_1, \ldots, x_n) \in G$ for which

$$\Sigma_{i=1}^n \Sigma_{j \in \mathbb{Z}} \|\langle T, x_i, r + j\rho\rangle\|_{\mathsf{C}} \leq n\rho^{-1}\|T\|_{\mathsf{C}}$$

has positive measure.

□

Lemma 5.4.2. *Suppose $T \in \mathcal{F}$. Suppose $D = \beta + (0,\rho)^n \subset \Omega$ is an open n-cube with center s, and $W = T\lfloor(D \times \mathbb{R}^N)$ has boundary $B = \partial W \in \mathcal{R}^{n-1}(\Omega \times \mathbb{R}^N)$ of finite mass. Let $p_D : D \times \mathbb{R}^N \to \{s\} \times \mathbb{R}^N$ be the projection $p_D(x,y) = (s,y)$, let $h_D : [0,1] \times D \times \mathbb{R}^N \to D \times \mathbb{R}^N$ be the homotopy $h_D(t,x,y) = (1-t)(x,y) + (tp(x), y)$, and let*

$$S_D = h_{D\#}([\![0,1]\!] \times B) + p_{D\#}(W).$$

Then

$$I\!M(S_D) \leq c\rho I\!M(B) + I\!M(W)$$

where c depends only on n, and $\partial S_D = \partial W$. In particular, spt $S_D \subset\subset \Omega \times \mathbb{R}^N$. Also, when $\mathcal{H}^{n-1}(\partial D \cap N_T) = 0$, $S_D \in \mathsf{cart}^*\ (D, \mathbb{R}^N)$ and

$$\|u_{S_D}\|_{L^1} \leq \rho\Sigma_{F \in \mathcal{P}} \|u_{T_F}\|_{L^1}.$$

Proof. Obvious.

$\square$

Lemmas 5.4.2 and 5.4.1 give a sequence of currents of uniformly bounded mass converging to T. Namely, for each k, define

$$S_k = \llbracket(\Omega \setminus \bigcup_{A\in\mathcal{T}_{\rho_k}} A) \times \{0\}\rrbracket + \Sigma_{A\in\mathcal{T}_{\rho_k}} S_A \in \mathsf{cart}(\Omega, \mathbb{R}^N).$$

Then $T - S_k = \partial(\llbracket[0,1]\rrbracket \times \Sigma_{A\in\mathcal{T}_{\rho_k}} h_{A\#}(T\lfloor A))$, and

$$I\!M(\llbracket[0,1]\rrbracket \times \Sigma_{A\in\mathcal{T}_{\rho_k}} h_{A\#}(T\lfloor A)) \le \rho_k I\!M(T) \to 0.$$

Also, $\partial S_k = 0$. Thus $S_k \overset{\mathsf{C}}{\rightharpoonup} T$. It remains only to show that in fact, $S_k \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$.

Lemma 5.4.3. *Assume that* $\mathsf{Cart}(G, \mathbb{R}^N) = \mathsf{cart}(G, \mathbb{R}^N)$ *holds for any* $G \subset \mathbb{R}^{n-1}$ *with smooth boundary. Let* $M(t,x,y) = (x, ty)$, *and with notation as in Lemma 5.4.2, let*

$$S'_D = h_{D\#}(\llbracket[0,1]\rrbracket \times B) + M_\#(\llbracket[0,1]\rrbracket \times (B + p_{D\#}(B))) + \llbracket(\Omega \setminus D) \times \{0\}\rrbracket$$

Then $S'_D \in \mathcal{C}$, *where* $\mathcal{C}$ *is the weak sequential closure of*

$$\{\llbracket G_u\rrbracket \mid u \in C^\infty(\Omega, \mathbb{R}^N) \textit{ and spt } u \subset D \setminus \partial D\}.$$

Proof. Let $\beta = (\beta_1, \dots \beta_n)$ be such that $D = \beta + [0,\rho]^n$. Choose $a \in (\beta_1, \beta_1 + \rho)$ such that $\langle B, x_1, a\rangle \in \mathcal{R}^{n-2}(\Omega, \mathbb{R}^N)$ and its mass is at most $\frac{1}{\rho} I\!M(B)$. Then take bi-Lipschitz orientation preserving homeomorphisms $f : \partial D \cap \{x_1 \le a\} \to [0, \frac{1}{2}] \times [0,1]^{n-2}$ and $g : \partial D \cap \{x_1 \ge a\} \to [\frac{1}{2}, 1] \times [0,1]^{n-2}$, such that on the set $L = \partial D \cap \{x_1 = a, x_2 = \beta_2\}$, the restrictions $f|_L$, $g|_L$ are equal, and are an affine bijection from L to $\{\frac{1}{2}\} \times [0,1]^{n-2}$. Let $b : C \to \partial D$ be equal to f^{-1} on $[0, \frac{1}{2}] \times [0,1]^{n-2}$ and to g^{-1} on $[\frac{1}{2}, 0] \times [0,1]^{n-2}$. Let

$$B' = (f_\#(B\lfloor(D \cap \{x_1 \le a\})) + g_\#(B\lfloor(D \cap \{x_1 \ge a\})))\lfloor(0,1)^{n-1} \times \mathbb{R}^N.$$

Then $B' \in \mathsf{cart}((0,1)^{n-1}, \mathbb{R}^N)$, $b_\#(B') = B$, and the mass of B' is finite.

Let i be inclusion of $(0,1)^{n-1} \times \mathbb{R}^{n-1}$ into $\mathbb{R}^{n-1} \times \mathbb{R}^N$. For $V \in \mathsf{cart}((0,1)^{n-1}, \mathbb{R}^N)$ with spt $V \subset (0,1)^{n-1} \times B_R(0)$ and such that $\partial V \in \mathcal{R}^{n-2}((0,1)^{n-2} \times \mathbb{R}^N)$ and $\partial b_\# V = 0$, define

$$\begin{aligned}\tilde{V} = h_\#(\llbracket[0,1]\rrbracket \times b_\#(V)) + M_\#(\llbracket[0,1]\rrbracket \times \partial h_\#(\llbracket[0,1]\rrbracket \times b_\#(V))) +& \\ \llbracket G_0\rrbracket\lfloor((\Omega \setminus D) \times \mathbb{R}^N)&.\end{aligned}$$

The aim now is to show that for any such V, $\tilde{V} \in \mathcal{C}$. For (V_k) a sequence of such currents converging C-weakly to V, it is clear that $\tilde{V}_k \overset{\mathsf{C}}{\rightharpoonup} \tilde{V}$. Let $M' : [0,1] \times \mathbb{R}^{n-1} \times \mathbb{R}^N \to \mathbb{R}^{n-1} \times \mathbb{R}^N$ be the map $M'(t,x,y) = (x, ty)$. Let $V' = V + M'(\llbracket[0,1]\rrbracket \times \partial(i_\# V)) + \llbracket(\mathbb{R}^n \setminus D) \times \{0\}\rrbracket$. Let $\psi_k : \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ be the dilation with fixed point s

and scaling factor $(1-\frac{1}{k})$. Then $\psi_{k\#}V'\lfloor D\times\mathbb{R}^N \rightharpoonup V'\lfloor D\times\mathbb{R}^N$, and $b_\# V' = b_\# V$. Thus it suffices to assume that spt $(V - [\![G_0]\!]) \subset\subset (0,1)^{n-1}\times\mathbb{R}^N$. In fact, it suffices to assume that V is given by integration over the graph of a smooth function with compact support. To see this, choose $\psi \in C_c^\infty(\Omega)$ such that $\psi|_{\pi(\text{spt }(V-[G_0]))} \equiv 1$ and spt $\psi \subset D$. Also choose $\phi \in C^\infty(\mathbb{R})$ such that $\phi((\infty,0]) = \{0\}$, $\phi|_{[0,R]} = \text{id}_{[0,R]}$ and $\phi([2R,\infty)) = \{\frac{3}{2}R\}$. let $\Psi(x,y) = (x, \psi(x)\phi(|y|)\frac{y}{|y|}$ for $y \neq 0$ and let $\Psi(x,0) = (x,0)$. While Ψ is not proper, $\Psi_\#$ can be defined as was p_k in Lemma 5.5.1. It is Lipschitz, and hence maps cartesian currents to cartesian currents continuously. Moreover, if $u \in C^\infty((0,1)^{n-1},\mathbb{R}^N)$, then $\Psi_\#[\![G_u]\!] = [\![G_{\tilde{u}}]\!]$, where

$$\tilde{u}(x) = \begin{cases} 0 & \text{if } u(x) = 0) \\ \psi(x)\phi(u(x))\frac{u(x)}{|u(x)|} & \text{otherwise} \end{cases}$$

which is a smooth, compactly supported function. Finally, $\Psi_\# V = V$. It follows by Lemma 5.1.10 that V lies in the weak sequential closure of

$$\{[\![G_u]\!] \mid u \in C_c^\infty((0,1)^{n-1},\mathbb{R}^N)\}$$

and thus it suffices to assume that V itself lies in this set. Say $V = [\![G_u]\!]$.

Let $q: (D\setminus\{s\}) \to \partial D$ be radial projection from s. For $x \in D\setminus\{s\}$, $u_{\tilde{V}}(x) = u(b^{-1}(q(x)))$. Thus $u_{\tilde{V}}|_{D\setminus\{s\}}$ is locally Lipschitz. Let $(\psi_k) \subset C^\infty(\mathbb{R})$ be a sequence of functions such that

- For each k, $\int_\Omega |D\psi_k| = 2$,
- each ψ_k is zero on $\mathbb{R}\setminus(\frac{1}{2k}, 1-\frac{1}{2k})$ and one on $(\frac{1}{k}, 1-\frac{1}{k})$.

Let $\Psi_k : \Omega \to \mathbb{R}$ be the function $\Psi_k(x) = \psi_k(\max_i|x_i - s_i|)$. Then each $\Psi_k u_{\tilde{V}}$ is equal to a Lipschitz function away from s, so by Lemma 5.1.9 $\{[\![G_{\Psi_k u_{\tilde{V}}}]\!]\} \subset \mathsf{Cart}(\Omega,\mathbb{R}^N)$. It easy to see that $[\![G_{\Psi_k u_{\tilde{V}}}]\!] \overset{\mathsf{C}}{\rightharpoonup} \tilde{V}$.

□

Corollary 5.4.4. $S_k \in \mathsf{Cart}(\Omega,\mathbb{R}^N)$

Proof. Let $\{A_1,\ldots\}$ be an enumeration of $\mathcal{T}_{\rho_k}$. With notation as in the above Lemma, let

$$S_l = [\![\Omega\setminus\bigcup_{i=1}^{l} A_i \times\{0\}]\!] + \Sigma_{i=1}^{l} S'_{A_i}$$

Suppose it has been shown that $S_l \in \mathsf{Cart}(\Omega,\mathbb{R}^N)$. By Lemma 5.1.10, with $V = \{u \in C^\infty(\Omega,\mathbb{R}^N) \mid \text{spt } u \cap \bigcup_{i=1}^{l} A_l = \emptyset\}$, and

$$F(T) = T\lfloor(\Omega\setminus\bigcup_{i=1}^{l} A_l)\times\mathbb{R}^N + S_l\lfloor(\bigcup_{i=1}^{l} A_l\times\mathbb{R}^N),$$

it follows that $S_{l+1} \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$. Since $S_l \overset{C}{\rightharpoonup} \Sigma_i S'_{A_i}$, $\Sigma_i S'_{A_i} \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$. Then by Lemma 5.2.3, since $S_k - \Sigma_i S'_{A_i}$ is supported over the centers of the cubes in $\mathcal{T}_{\rho_k}$, it follows that $S_k \in \mathsf{Cart}(\Omega, \mathbb{R}^N)$.

□

Theorem 5.4.5. $\mathsf{Cart}(\Omega, \mathbb{R}^N) = \mathsf{cart}(\Omega, \mathbb{R}^N)$

Proof. This follows immediately from the above corollary and the fact that $S_k \overset{C}{\rightharpoonup} T$.

□

5.5 Approximating currents by compactly supported currents

Lemma 5.5.1. *Suppose* $T \in \mathsf{cart}(\Omega, \mathbb{R}^N)$. *Then there is a sequence* $(Z_k) \subset \mathsf{cart}(\Omega, \mathbb{R}^N)$ *such that* $Z_k \overset{C}{\rightharpoonup} T$ *as* $k \to \infty$ *and spt* $Z_k \subset \Omega \times B_{R_k}(0)$ *for some* R_k.

Proof. For $k \in \mathbb{N}$, let $p_k : \mathbb{R}^N \to \overline{B_k(0)}$ be the following retraction map:

$$p_k(y) = \begin{cases} k\frac{y}{|y|} & \text{if } |y| \geq k \\ y & \text{if } |y| \leq k \end{cases}$$

Let $\pi_k = \mathrm{id}_\Omega \times p_k$. Note that this map has Lipschitz constant 1, and for any $s > 0$, $\pi_k|_{\Omega \times \overline{B_s(0)}}$ is proper. Let $s : \Omega \times \mathbb{R}^N \to \mathbb{R}$ be the map $s(x, y) = |y|$. This map also has Lipschitz constant 1.

For any $R > 0$, let $X_R = T \lfloor \Omega \times B_R(0)$. Let $Y_R = \langle T, s, R \rangle$ for each R for which this is well defined. By [48, Lemma 28.5], this is the case for almost all $R > 0$, and $Y_R = \partial X_R$. Also

$$\int_0^\infty \mathbb{M}(Y_R) dT \leq \mathbb{M}(T)$$

so there is a sequence $R_k \uparrow \infty$ for which $\mathbb{M}(Y_{R_k}) \to 0$. Since spt X_R and spt Y_R lie in $\Omega \times \overline{B_R(0)}$, the push-forwards $\pi_{k\#}(X_{R_l})$ and $\pi_{k\#}(Y_{R_l})$ are well defined for all k. They satisfy the mass estimates

$$\mathbb{M}(\pi_{k\#}(X_{R_l})) \leq \mathbb{M}(T)$$

$$\mathbb{M}(\partial \pi_{k\#}(X_{R_l})) = \mathbb{M}(\pi_{k\#}(Y_{R_l})) \leq \mathbb{M}(Y_{R_l}) \to 0$$

Applying the compactness Theorem for integer multiplicity rectifiable currents, for any $k \in \mathbb{N}$ there is some $Z_k \in \mathcal{R}^n(\Omega \times \mathbb{R}^N)$ which is a weak limit point for the sequence $(\pi_{k\#}(X_{R_l}))$. From the weak lowersemicontinuity of the mass norm, $\mathbb{M}(Z_k) \leq \mathbb{M}(T)$ and $\partial Z_k = 0$.

Given a differentiable n-form $\omega \in \mathcal{D}^n(\Omega \times \mathbb{R}^N)$, spt $\omega \subset \Omega \times B_k(0)$ for all sufficiently large $k \in \mathbb{N}$. For any such k, and for $R > k$, $X_R(\omega) = T(\omega)$, and since $\pi_k|_{\Omega \times B_k(0)} = \mathrm{id}_{\Omega \times B_k(0)}$, it follows that $\pi_{k\#}(X_R)(\omega) = T(\omega)$. From the definition of weak convergence, it follows that $Z_k(\omega) = T(\omega)$. Thus it follows that $Z_k \rightharpoonup T$.

Since $\pi \circ \pi_k = \pi$, $Z_k \in \mathsf{cart}^* (\Omega, \mathbb{R}^N)$. Since for almost all x, $u_{Z_k}(x) = \max(k, u_T)$, $Z_k \in \mathsf{cart}(\Omega, \mathbb{R}^N)$. □

Next, it is necessary to approximate vertical boundaryless currents by vertical boundaryless currents with compact support.

Lemma 5.5.2. *Let $\Omega \subset \mathbb{R}^n$ be a bounded open set with smooth boundary. Let $T \in$ $\mathsf{cart}(\Omega, \mathbb{R}^N)$ be a current whose support lies in $\Omega \times B_R(0)$, where R is some positive number. Then there is a sequence of currents $(W_k) \subset \mathsf{cart}(\Omega, \mathbb{R}^N)$ such that $W_k \overset{C}{\rightharpoondown} T$, and $\pi(spt\ (W_k - [\![G_0]\!])) \subset\subset \Omega \times \mathbb{R}^N$*

Proof. Let $D : \Omega \times \mathbb{R}^N \to \mathbb{R}$ be the function $D(x, y) = \operatorname{dist}(x, \partial\Omega)$. For $\delta > 0$, let $\Omega_\delta = D^{-1}[(\delta, \infty)]$. Fix $\epsilon > 0$ so that D is smooth and Lipschitz on $G = \Omega \setminus \overline{\Omega_\epsilon}$. Then the nearest point projection map $p : G \to \partial\Omega \times \mathbb{R}^N$ is well defined, Lipschitz and smooth. For $\delta \in (0, \epsilon)$ let $p_\delta : G \to \partial\Omega_\delta \times \mathbb{R}^N$ be the map

$$p_\delta(x, y) = \left(p(x) + \delta \frac{x - p(x)}{|x - p(x)|}, y\right)$$

Then p_δ is nearest point projection onto $\partial\Omega_\delta$, satisfying a Lipschitz bound which is independent of δ. Let $q_\delta : G \times \mathbb{R}^N \to \Omega \times \mathbb{R}^N \setminus (\Omega_\delta \times \mathbb{R}^N)$ be the map

$$q_\delta(x, y) = \begin{cases} (x, y) & \text{if } x \notin \Omega_\delta \\ p_\delta(x, y) & \text{if } x \in \Omega_\delta \end{cases}$$

Again, this satisfies a Lipschitz constant which is independent of δ.

Let $H(x, y, t) = (x, ty)$. Let $B = H_\#([\![0, 1]\!] \times T)$. Then $\partial B = T$ and spt $B \subset \Omega \times B_R(0)$. Applying [48, Lemma 28.5] to the Lipschitz functions D and $-D$, there is a full-measure set C of $\delta \in (0, \epsilon)$ for which $T_\delta = \langle B, D, \delta \rangle$ is defined, has finite mass, and satisfies the equality

$$T_\delta = \partial(B \lfloor \Omega_\delta \times \mathbb{R}^N) - T \lfloor \Omega_\delta \times \mathbb{R}^N$$

Such T_δ is vertical, for it is supported in $\partial\Omega_\delta \times \mathbb{R}^N$. Moreover,

$$\partial T_\delta = -\partial(T \lfloor \Omega_\delta \times \mathbb{R}^N) = \partial(T \lfloor \Omega \times \mathbb{R}^N \setminus \overline{\Omega_\delta \times \mathbb{R}^N})$$

and so $T \lfloor \Omega_\delta \times \mathbb{R}^N - T_\delta$ and $T \lfloor \Omega \times \mathbb{R}^N \setminus \overline{\Omega_\delta \times \mathbb{R}^N} + T_\delta$ are boundaryless and have finite mass.

Choose δ and a sequence $\delta_k \to 0$ from C. Let

$$N_k = q_{\delta_k \#}(T \lfloor \Omega \times \mathbb{R}^N \setminus \overline{\Omega_\delta \times \mathbb{R}^N} - T_\delta).$$

This is a boundaryless current whose mass is bounded independently of k. Also, note that $q_{\delta_k}|_{\Omega \times \mathbb{R}^N \setminus \overline{\Omega_{\delta_k} \times \mathbb{R}^N}}$ is the identity map, and $q_{\delta_k}[G \cap \Omega_{\delta_k}] \subset \partial\Omega_{\delta_k}$. Hence $N_k \lfloor (\Omega \setminus \overline{\Omega_{\delta_k}}) \times \mathbb{R}^N = T \lfloor (\Omega \setminus \overline{\Omega_{\delta_k}}) \times \mathbb{R}^N$, and spt $N_k \subset (\Omega \setminus \overline{\Omega_{\delta_k}}) \times \mathbb{R}^N$, so $N_k =$

$N_k \llcorner (\Omega \setminus \overline{\Omega_{\delta_k}}) \times \mathbb{R}^N + N_k \llcorner \partial\Omega_{\delta_k} \times \mathbb{R}^N$. Applying the boundary operator to both sides of this equation gives

$$\begin{aligned} 0 &= \partial(T \llcorner \Omega \times \mathbb{R}^N \setminus \overline{\Omega_\delta \times \mathbb{R}^N}) + \partial(N_k \llcorner \partial\Omega_{\delta_k} \times \mathbb{R}^N) \\ &= -\partial(T \llcorner \Omega_{\delta_k} \times \mathbb{R}^N) + \partial(N_k \llcorner \partial\Omega_{\delta_k} \times \mathbb{R}^N) \end{aligned}$$

So defining $W_k = T \llcorner \Omega_{\delta_k} \times \mathbb{R}^N - N_k \llcorner \partial\Omega_{\delta_k} \times \mathbb{R}^N + [\![\Omega \setminus \Omega_{\delta_k} \times \{0\}]\!]$ gives a sequence which satisfies the requirements of the Theorem.

□

Bibliography

[1] S. Altschul, R. Bundschuh, R. Olsen, and T. Hwa. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Research*, 29(2):351–361, 2001.

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

[3] A. Bansal and T. Meyer. Evolutionary analysis by whole-genome comparisons. *Journal of Bacteriology*, 184(8):2260–2272, April 2002.

[4] S. Batzoglou, L. Pachter, J. P. Mesirov, B. Berger, and E. S. Lander. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research*, 7(10):950–8, July 2000.

[5] D. Benson, D. Lipman, and J. Ostell. Genbank. *Nucleic Acids Research*, 13(21):2963–2965, 1993.

[6] C. Brown, B. Hendrich, J. Rupert, R. Lafreniere, Y. Xing., J. Lawrence, and H. Willard. The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71:527–542, October 1992.

[7] J. Brown. The Ribonuclease P Database. *Nucleic Acids Research*, 27:314, 1999.

[8] C. Burge. *Identification of genes in human genomic DNA*. PhD thesis, Stanford University, March 1997.

[9] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1):78–94, April 1997.

[10] R. Carter, I. Dubchak, and S. Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Research*, 19(29):3928–38, October 2001.

[11] The International Human Genome Sequencing Consortium. A physical map of the human genome. *Nature*, 409:934–941, Febuary 2001.

[12] P. Douglas, C. Wu, and P Burbelo. Human rhogap domain-containing proteins: structure, function and evolutionary relationships. *Federation of European Biochemical Societies Letters*, 528(1-3):27–34, September 2002.

[13] T. Down and T. Hubbard. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Research*, 3(12):458–61, March 2002.

[14] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, 1998.

[15] S. Eddy. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–29, December 2001.

[16] S. Gregory *et al.* A physical map of the mouse genome. *Nature*, 418:743–750, August 2002.

[17] H. Federer. *Geometric Measure Theory*. Springer, 1969.

[18] G. Fox and C. Woese. 5S RNA secondary structure. *Nature*, 256:505–507, 1975.

[19] M. Gelfand, A. Mironov, and P. Pevzner. Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Science*, 93:9061–9066, 1996.

[20] M. Giaquinta, G. Modica, and J. Souček. Cartesian currents and variational problems for mappings into spheres. *Ann. Scuola Norm. Sup. Pisa*, 16:393–485, 1989.

[21] M. Giaquinta, G. Modica, and J. Souček. Cartesian currents, weak diffeomorphisms and nonlinear elasticity. *Archive for Rat. Mech. Anal.*, 106:97–159, 1989.

[22] M. Giaquinta, G. Modica, and J. Souček. Erratum and addendum to "Cartesian currents, weak diffeomorphisms and nonlinear elasticity". *Archive for Rat. Mech. Anal.*, 109:385–392, 1990.

[23] K. Giles, J. Daly, D. Beveridge, A. Thomson, D. Voon, H. Furneaux, J. Jazayeri, and P. Leedman. The 3'-UTR of p21WAF1 mRNA is a composite *cis*-acting sequence bound by RNA-binding proteins from breast cancer cells, including HuR and poly(C)-binding protein (CP1). *Journal of Biological Chemistry*, November 2002.

[24] W. Gish, 1996-2000. `http://blast.wustl.edu`.

[25] J. Gorodkin, B. Knudsen, C. Zwieb, and T. Samuelsson. SRPDB (signal recognition particle database). *Nucleic Acids Research*, 29(1):169–170, 2001.

[26] D. Higgins, J. Thompson, and T. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

[27] I. Hofacker, M. Fekete, and P. Stadler. Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*, 319:1059–1066, June 2002.

[28] I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, 125:167–188, 1994.

[29] Y.-K. Hong, S. Ontiveros, and W. Strauss. A revision of the human XIST gene organization and structural comparison with mouse xist. *Mammalian Genome*, 11(3):220–4, March 2000.

[30] B. James, G. Olsen, J. Liu, and N. Pace. The secondary structure of ribonuclease P RNA, the catalytic element of a ribonucleoprotein enzyme. *Cell*, 52:19–26, 1988.

[31] R.-P. Jansen. mRNA localization: Message on the move. *Nature Reviews Molecular Cell Biology*, 2:247–256, 2001.

[32] H. Koch, K. Hofmann, and N. Brose. Definition of munc13-homology-domains and characterization of a novel ubiquitously expressed munc13 isoform. *Journal of Biochemistry*, 1(349(Pt 1)):247–53, July 2000.

[33] Y. Lee, R. Sultana, G. Pertea, J. Cho, S. Karamycheva, J. Tsai, B. Parvizi, F. Cheung, V. Antonescu, J. White, I. Holt, F. Liang, and J. Quackenbush. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Research*, 12(3):493–502, March 2002.

[34] Lodish and *et al. Molecular Cell Biology.* W.H. Freeman, NY, 2000.

[35] J. Malý. l^p-approximation of jacobians. *Comment. Math. Univ. Carolinae*, 4(32):659–666, 1991.

[36] D. Mathews and D. Turner. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, 317(19):191–203, 2002.

[37] Q. Meng and R. Switzer. cis-acting sequences of bacillus subtilis pyrg mrna essential for regulation by antitermination. *Journal of Bacteriology*, 184(23):6734–8, December 2002.

[38] NCBI. `http://www.ncbi.nlm.nih.gov/BLAST/`.

[39] NCBI. `http://www.ncbi.nlm.nih.gov/Entrez/`.

[40] NCBI. `http://www.ncbi.nlm.nih.gov/UniGene/`.

[41] A. Nekrutenko, K. Makova, and W.-H. Li. The KA/KS ratio test for assessing the protein-coding potential of genomic regions: An empirical and simulation study. *Genome Research*, 12:198–202, 2002.

[42] R. Nussinov, G. Piecznik, J. Griggs, and D. Kleitman. Algorithms for loop matchings. *SIAM Journal of Applied Mathematics*, 35(1):68–82, 1978.

[43] N. Pace, D. Smith, G. Olsen, and B. James. Phylogenetic comparative analysis and the secondary structure of ribonuclease—a review. *Gene*, 82:65–75, 1989.

[44] M. Regalia, M. Rosenblad, and T. Samuelsson. Prediction of signal recognition particle RNA genes. *Nucleic Acids Research*, 30(15):3368–3377, 2002.

[45] M. Rhoades, B. Reinhart, L. Lim, C. Burge, B. Bartel, and D. Bartel. Prediction of plant microRNA targets. *Cell*, 110(4):513–20, August 2002.

[46] E. Rivas and S. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 7(16):583–605, July 2000.

[47] E. Rivas and S. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *Biomed Central Informatics*, 2(8), 2001.

[48] L. Simon. *Lectures on geometric measure theory.* Centre for Mathematical Analysis, Canberra, 1984.

[49] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195197, 1981.

[50] V. Solovyev, A. Salamov, and C. Lawrence. Identification of human gene structure using linear discriminant functions and dynamic programming. *Proceedings of the International Conference on Intelligent Systems in Molecular Biology*, 12:367–75, 1995.

[51] T. Spies, M. Bresnahan, and J. Strominger. Human major histocompatibility complex contains a minimum of 19 genes between the complement cluster and HLA-B. *Proceedings of the National Academy of Sciences*, 86(22):8955–8, November 1989.

[52] S. Twigger, J. Lu, M. Shimoyama, D. Chen, D. Pasko, H. Long, J. Ginster, C. Chen, R. Nigam, A. Kwitek, J Eppig, L. Maltais, D. Maglott, G. Schuler, H. Jacob, and P. tonellato. Rat genome database (RGD): mapping disease onto the genome. *Nucleic Acids Research*, 30(1):125–8, January 2002.

[53] A Walter, D Turner, J Kim, M Lyttle, P Muller, D Mathews, and M. Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proceedings of the National Academy of Sciences*, 91:9218–9222, 1994.

[54] R.-F. Yeh, L. P. Lim, and C Burge. Computational inference of homologous gene structures in the human genome. *Genome Research*, 11:803–816, 2001.

[55] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–48, January 1981.