Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA

Eva K. Freyhult,[1] Jonathan P. Bollback,[2] and Paul P. Gardner[3,4]

[1]*The Linnaeus Centre for Bioinformatics, Uppsala University, 75124 Uppsala, Sweden;* [2]*Evolution Department, Biological Institute, University of Copenhagen, 2100 Copenhagen, Denmark;* [3]*Molecular Evolution Group, Institute of Molecular Biology and Physiology, University of Copenhagen, 2100 Copenhagen, Denmark*

Homology search is one of the most ubiquitous bioinformatic tasks, yet it is unknown how effective the currently available tools are for identifying noncoding RNAs (ncRNAs). In this work, we use reliable ncRNA data sets to assess the effectiveness of methods such as BLAST, FASTA, HMMer, and Infernal. Surprisingly, the most popular homology search methods are often the least accurate. As a result, many studies have used inappropriate tools for their analyses. On the basis of our results, we suggest homology search strategies using the currently available tools and some directions for future development.

[Supplemental material is available online at www.genome.org and http://www.binf.ku.dk/~pgardner/bralibase/bralibase3/.]

Compared with the relatively trivial task of protein homology search, ncRNA homology search is more challenging because of the fact that intra- and intermolecular base pairs are, in evolutionary terms, preserved to a higher degree than the sequence. The wobble GU and other noncanonical base pairs allow RNA sequences to evolve seemingly unrelated sequences along nearly neutral paths through structure space (e.g., $A \cdot U \leftrightarrow G \cdot U \leftrightarrow G \cdot C$). Thus, specialized homology search techniques, such as nucleotide specific scoring schemes (States et al. 1991), profile hidden Markov model (profile HMMs) (Haussler et al. 1993; Krogh et al. 1994), and covariance models (CMs) (Eddy and Durbin 1994), are necessary for accurate ncRNA homology search.

The goal of this study is to identify programs that balance sensitivity (true predictions) and specificity (false predictions) for practical ncRNA homology search situations. We use large high-quality ncRNA data sets and randomized control data sets to test the 12 homology search programs summarized in Table 1. Briefly, sequences are sampled from each ncRNA data set and then used as input sequences for each algorithm against the original (true homologs) and randomized data sets. Our test data sets are composed of a subclass of ncRNAs that tend to be highly structured, and therefore there is more information for homology detection than for unconstrained ncRNAs. The algorithms that do not perform well on these data sets are not likely to perform better on more challenging classes of ncRNAs. To ensure our results reflect practical scenarios, we have used both predicted alignments and secondary structures to generate input data for the alignment and structure-based methods.

Homology search programs fall into one of three classes: sequence based methods, profile HMM methods, and structure enhanced methods (Fig. 1). In addition to evaluating homology search programs, we extend the use of ancestral sequence reconstructions (ASR) and introduce the novel phylogeny-based predictive sequence reconstruction (PSR) method for use in homology searches (Collins et al. 2003; McCormack 2003; Qian and Goldstein 2003; Cai et al. 2004) to the RNA homology search problem (see Supplemental Fig. 1). Briefly, we discuss each of these in turn.

The most popular homology search methods are sequence based. The local matching of two sequences has been solved by Smith and Waterman (1981) in a mathematically optimal fashion using a dynamic programming procedure. However, this method is too slow for most practical homology search situations, where the database length is large. Hence, heuristic methods such as BLAST and FASTA, which speed the search procedure but at a cost to accuracy, are often used.

Profile HMMs have been used for detecting patterns in multiple sequences (Haussler et al. 1993; Krogh et al. 1994); assessments of profile HMMs on protein data sets have proven that these are more accurate than sequence methods alone (Brenner et al. 1998; Park et al. 1998; Lindahl and Elofsson 2000; Madera and Gough 2002). The basic usage of a profile HMM is to convert an input alignment into a probabilistic model, which is used to scan a database for homologous sequences. The fundamental concept of profile HMMs can be understood by considering nucleotide frequencies in each column of an alignment. In the absence of gaps, the probability that a given sequence is generated by the same evolutionary processes as those in the alignment can be estimated by the product of position specific nucleotide frequencies. The architecture proposed by Krogh et al. (1994) (see Supplemental Fig. 2) allows for insertions and deletions in the model, and, in addition, deletions can be modeled in a position-dependent manner. To account for overrepresented sequences in the input alignment, tree-weighting schemes can be used (Durbin et al. 1998), and there are schemes to avoiding over-fitting and to account for unobserved data in the input (Sjölander et al. 1996).

[4]Corresponding author.
E-mail PPGardner@bi.ku.dk; fax 45-35321300.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5890907.

Table 1. Program descriptions, URLs, and references for each of the 12 programs used in this study

Program	Description and URL	Version	Reference
Sequence-based methods			
NCBI-BLAST WU-BLAST	The query database is tabulated as short sequences (seeds), which are scanned during the initial search phase; short matches are subsequently extended and scored. URLS: NCBI, www.ncbi.nlm.nih.gov/BLAST WU, blast.wustl.edu	NCBI: 2.2.10 WU: 2.0	(Altschul et al. 1990), (Gish 2005)
FASTA	FASTA employs a lookup table to identify all matching words of length ktup. Diagonals of mutually supporting matches are located, linked, and extended. High-scoring matches are finally realigned using a local banded Smith-Waterman algorithm (Chao et al. 1992). URL: fasta.bioch.virginia.edu	3.4	(Pearson and Lipman 1988)
ParAlign	A parallel computing technology, SIMD (Single Instruction Multiple Data), is used to compute exact ungapped alignments. A novel heuristic is used to compute gapped alignments for high scoring ungapped hits. The highest scoring database matches are realigned using a rigorous SIMD-based Smith-Waterman algorithm. URL: www.paralign.org	3.4.3	(Saebø et al. 2005)
SSEARCH	Implements the Smith-Waterman local alignment algorithm (Smith and Waterman 1981). URL: fasta.bioch.virginia.edu	3.4	(Pearson 1991)
Profile HMM methods			
HMMer	A profile HMM approach with a novel "Plan 7" architecture that distinguishes between global and local alignments probabilistically and excludes transitions from insert to delete states and vice versa. URL: hmmer.wustl.edu	1.8.4 & 2.3.2	(Eddy 1998)
SAM	This package uses the original profile HMM architecture (Krogh et al. 1994) discussed in the text and displayed in Supplemental Figure 2. URL: www.cse.ucsc.edu/research/compbio/sam.html	3.5	(Hughey and Krogh 1996), (Karplus et al. 1998)
Structure-based methods			
ERPIN	An input alignment with structure annotation is converted into a combination of single sequence and helical lod-score based weight matrices. These profiles can then be used to rapidly screen a database for matching helical profiles and classical dynamic programming for the alignment of single-stranded regions. URL: tagc.univ-mrs.fr/erpin	4.2.5	(Gautheret and Lambert 2001)
Infernal	Implements a covariance model (CM) as discussed in the text and illustrated in Supplemental Figure 2. Additional features added during this investigation are an "effective sequence number" weighting scheme and Dirichlet mixture priors (Sjölander et al. 1996). URL: infernal.janelia.org	0.7	(Eddy 2002)
RaveNnA	Converts a CM generated by Infernal into a profile HMM. This is used to rapidly filter the database for high-scoring matches, which can be aligned using the slower but more accurate Infernal package. URL: bio.cs.washington.edu/supplements/zasha-ravenna	0.2f	(Weinberg and Ruzzo 2006)
RSEARCH	Implements a CM for a single input sequence and structure. BLOSUM-like score matrices (Henikoff and Henikoff 1992) called RIBOSUM matrices are used to score database sequence matches to helical or single-stranded regions of the query. URL: http://selab.janelia.org/software.html#rsearch	1.1	(Klein and Eddy 2003)
RSmatch	Input and database sequences are folded using RNAfold (Hofacker et al. 1994) (or similar). The structures are decomposed into subcomponents, which are organized into a tree model, and the database is screened for significant hits using a tree alignment procedure. The alignment is scored using a combination of base-pair and single-strand score matrices. URL: exon.umdnj.edu/software/RSmatch	1.2	(Liu et al. 2005)

Structure-enhanced methods are frequently based on CMs, which are an analog of profile HMMs that include pairwise interactions due to RNA secondary structure. Whereas profile HMMs consist of a linear HMM architecture suitable for modeling linear protein sequences, tree-like CMs model tree-like RNA secondary structures that allow for base-pairing interactions. States within the CM capture paired and unpaired regions while allowing insertions and deletions. To picture this, imagine the profile HMM model in Supplemental Figure 2 with base pairs between distant sites. Several new states need to be added to the model to accommodate this more complex structure. In the paired sites, deletions now include either a single 5′ or 3′ base or the entire base pair, and insertions can now be between either the 5′ or 3′ ends of a base pair. Bifurcation states are also included in the CM to allow for multiloops. The basic CM search procedure is analogous to the use of profile HMMs. An alignment replete with a structure annotation is provided by the user; this is used to train a CM that is specific to the input data, which can then be used to search a query database (Eddy and Durbin 1994; Durbin et al. 1998).

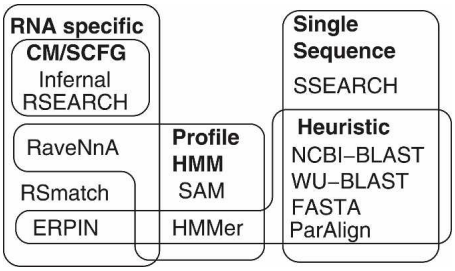

Figure 1. An overview of homology search methods. A Venn diagram illustrating an overview of the methods used in this study. Different methods are classified as heuristic, single sequence, profile HMM, stochastic context-free grammar (SCFG), and/or RNA specific.

Nearly all of the current homology search algorithms, with the exception of some profile HMM tree-weighting schemes (Durbin et al. 1998), ignore important evolutionary information contained in the underlying phylogenetic relationships among the input sequences. For example, the branching order and distance between sequences is largely ignored. To address these shortcomings and aid the phylogenetically naive algorithms, we employ two probabilistic phylogenetic approaches: ASR and PSR. Briefly, each of these methods sample high-probability ancestral sequences that are added to the query sequences. In this way, additional information derived from the phylogenetic relationships is added to the search, in theory boosting remote homolog detection.

Caveats to algorithm assessments

There are several limitations to any algorithm assessment. We outline the most important issues below.

Test data sets

We take for granted the accuracy of structural alignments taken from the literature, many of which have been constructed using the programs we are studying. However, given this limitation, the analysis of a large and diverse data set should outweigh any possible errors due to data set inaccuracies. We make one final point regarding the nature of our test data sets; the data sets used here are conserved "ideal" ncRNAs and may not be representative of other ncRNAs. Although it is likely that other families of ncRNAs will be less conserved and less "ideal," it seems clear that if a method fails under "ideal" circumstances, its performance is unlikely to improve in more challenging circumstances.

Tool abuse

Frequently, researchers may apply a tool to a task for which it is not designed. For example, in this study we have applied sequence-based tools to structured ncRNAs, assuming that sites are independent. This is a common but poor assumption.

Tools improve

Many of the tools tested here are recent developments and are still under active development. Hence, not all observations will remain reproducible. In fact, we hope this study helps improve future performance.

Parameter settings

The performance of some of the programs may benefit from optimizing program parameters. Here, we have attempted to capture the essential features of each algorithm by using as many parameter combinations as was practical.

During the course of this investigation, we contacted the authors of each of the programs included in this study (see Acknowledgments). We provided access to the data sets, scripts, and a preprint of the article for the authors from the BRaliBase Web site (www.binf.ku.dk/~pgardner/bralibase). We found the comments we received invaluable for minimizing the costs of the above caveats.

Results

The following discussion contains a detailed summary of the results presented in Figures 2–4 and the Supplemental Tables 1–4 and Supplemental Figures 5–7. We begin by outlining the results

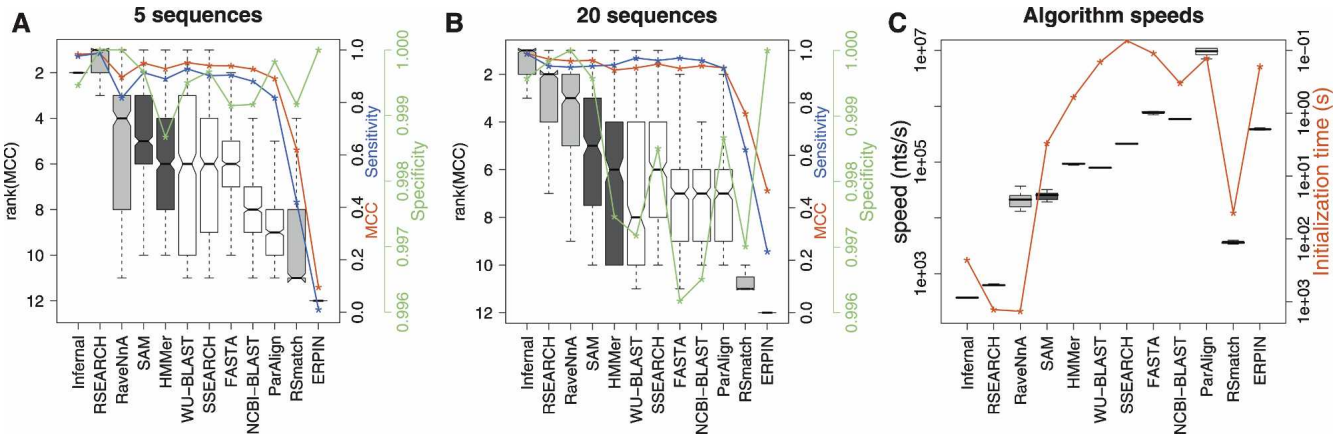

Figure 2. A comparison of the accuracy and efficiencies of homology search methods showing only the highest-ranking parameter settings for each algorithm from Supplemental Table 1. These were NCBI-BLAST (W7, 65%), WU-BLAST (W3), FASTA, ParAlign (65%), SSEARCH, HMMer (2.3.2, local), SAM (3.5, local), ERPIN, Infernal (0.7, local), RaveNnA, RSEARCH, and RSmatch. (*A*,*B*) Boxplots of algorithm ranks for the 5 and 20 sequence subsets, respectively. The blue curves show the median sensitivity, the green curve the median specificity, and the red curve the median MCC for each of the 12 programs. These accuracy values were computed by sampling either 5 or 20 sequences from the reference databases; these were used as input(s) to each algorithm for screening both the reference and a shuffled database. (*C*) Boxplots of algorithm speeds in nucleotides per second. The red curve shows median initialization times for the different programs. The single sequence, profile HMM, and RNA methods are displayed in unshaded, dark shaded, and lightly shaded boxes, respectively.

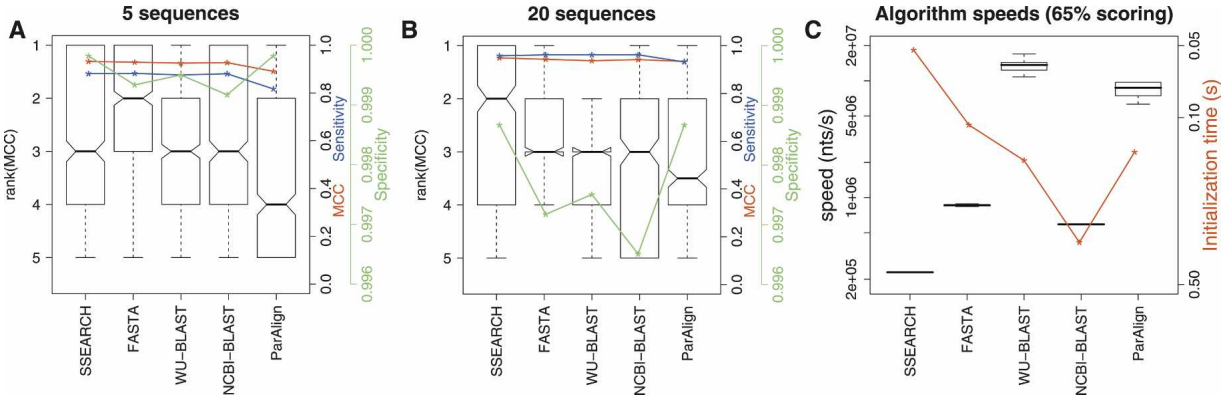

Figure 3. A comparison of the accuracy of sequence-based methods with the 65% scoring scheme and identical scoring parameters. These boxplots show the distributions of the ranks on MCC and timing data for each of the homology search methods when using a scoring scheme optimized for nucleotide sequences with 65% identity (match = +5, mismatch = −4, gapopen = 10, gapextension = 10). (*A*,*B*) Boxplots of algorithm ranks for the 5 and 20 sequence subsets, respectively. The blue curves show the median sensitivity, the green curve the median specificity, and the red curve the median MCC for each of the 12 programs. (*C*) Boxplots of algorithm speeds in nucleotides per second. The red curve shows median initialization times for the different programs.

for each method with the parameter settings that had the optimal ranking. Unless stated otherwise, we focus on the results for the smaller query subsets with just 5 sequences because the results do not differ significantly from the analysis using larger query subsets (20 sequences). Secondly, we outline the results for our secondary tests, which include a comparison of the sequence-based methods (NCBI-BLAST, WU-BLAST, FASTA, ParAlign, and SSEARCH) with identical scoring parameters. These scoring parameters are optimized for sequence identities ranging from 65%–100% (States et al. 1991) and are referred to as the 65% scoring scheme throughout this section. The other secondary tests we present are of RNA-centric scoring schemes for sequence-based methods, the application of phylogenetic sequence reconstruction to homology searching, and a scan of a section of the human genome.

Sequence-based searches

The accuracies of WU-BLAST and NCBI-BLAST were unsurprisingly similar when similar parameters were used, yet WU-BLAST was significantly faster (Fig. 3). The default scoring scheme used for NCBI-BLAST is tailored for sequences with 99% sequence homology, whereas WU-BLAST defaults are tailored for sequences with 65% sequence homology (States et al. 1991), which is more appropriate for our diverse ncRNA data sets (see Supplemental Tables 1 and 2). WU-BLAST has a more diverse array of options, including allowing a minimum seed length of 3 (W3) (compared with 7 [W7] for NCBI-BLAST). Hence, the parameter settings producing the best accuracy for WU-BLAST are not implemented in NCBI-BLAST. However, shorter seed lengths did come at a significant cost to program speed.

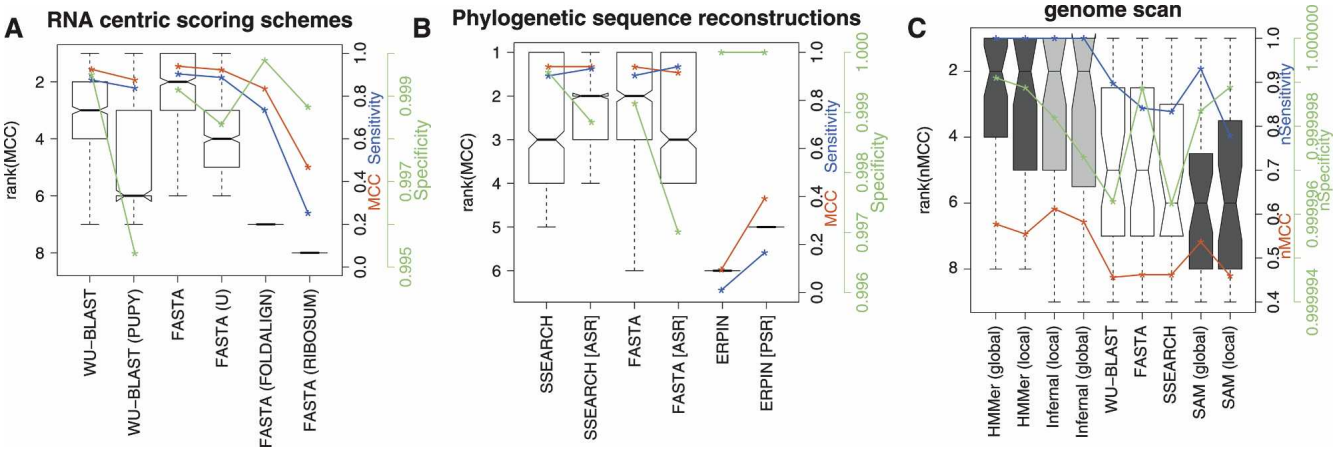

Figure 4. A comparison of the accuracy of methods using RNA-centric scoring matrices, phylogenetic sequence reconstructions, and the genome scan results. (*A*) A comparison of the accuracy of sequence-based methods with score matrices optimized for ncRNA. These boxplots show the distributions of the ranks on MCC for each of the homology search methods when using one of WU-BLAST (W7), WU-BLAST (W7, PUPY), FASTA, FASTA (*U*), FASTA (RIBOSUM), or FASTA (FOLDALIGN). These matrices are discussed in more detail in the text. (*B*) A comparison of FASTA, SSEARCH, and ERPIN with and without phylogenetic sequence reconstructions included in the input. Ancestral sequence reconstruction was used in the case of FASTA and SSEARCH and posterior predictive sequences in the case of ERPIN. Both *A* and *B* show results using 5 query sequences. (*C*) A set of representative programs from each category were run on human chromosome 12 (coordinates 90,000,000–130,000,000; ver NCBI35). The boxplot displays algorithm ranks; additionally, median nMCC, median nSensitivity, and median nSpecificity for each algorithm are displayed using the *y*-axis on the *right*. The single sequence, profile HMM, and RNA methods are displayed in unshaded, dark shaded, and lightly shaded boxes, respectively.

A comparison of FASTA and WU-BLAST (W3) is complicated given that the median ranking of WU-BLAST (W3) was higher than that of FASTA for the 5 sequence input data, but the lower quartile of the WU-BLAST (W3) ranks was much lower than that of FASTA (Fig. 3). Hence, most of the time WU-BLAST (W3) outperformed FASTA (the rankings reversed on the 20 sequence data sets). Yet, FASTA was significantly faster than WU-BLAST (W3) and compared well with NCBI-BLAST (W7, 65%) in terms of speed.

ParAlign was the fastest of the homology search tools in this study. However, ParAlign had low sensitivity compared with both FASTA and BLAST; this was true also for the 65% scoring test (see Methods and Fig. 3 for results).

SSEARCH generally outperformed the other sequence-based methods in terms of accuracy. However, SSEARCH performance was very closely correlated with WU-BLAST (W3), but WU-BLAST (W3) was significantly slower (see Supplemental Fig. 9). This observation was surprising given that SSEARCH employs no heuristics to improve speed whereas WU-BLAST (W3) demands seeds matching of at least three consecutive nucleotide positions; one would have expected the opposite to the results presented here.

Profile HMMs

The profile HMM programs evaluated here, SAM and HMMer, always outperformed the sequence-based methods. SAM usually outperformed HMMer in terms of accuracy, yet HMMer was significantly faster (Fig. 2). The results for HMMer show that version 2.3.2 is slightly better than v.1.8.4; this is in contrast to the HMMer documentation, which suggests the opposite for nucleotide sequences (because of protein-specific optimization). Earlier results based on protein data sets (Madera and Gough 2002) showed that profile searches could be improved by using SAM models and HMMer searches. We observed no such improvement on our ncRNA data sets (see Supplemental Tables 1 and 2).

Structure-enhanced homology search

The CM-based methods Infernal and RSEARCH both performed extremely well on these ncRNA data sets, providing predictions with very high sensitivity and specificity. These methods generally ranked either first or second in terms of the Matthews correlations coefficient (MCC) (see Methods for a definition) for every search. However, there was a significant cost in terms of CPU: Both take ~1 sec to search a kilobase using a 900-MHz processor. This is about 2 orders of magnitude slower than the profile HMM and sequence-based methods.

The Infernal package was upgraded during the course of this study to version 0.7. Sean Eddy and collaborators added Dirichlet mixtures (Sjölander et al. 1996) and effective sequence number scalings to the algorithm, which resulted in a significant performance boost for both the 5 and 20 sequence data sets (see Supplemental Tables 1 and 2).

ERPIN predictions are generally very conservative, especially for the small data set or when sequence identity is high (frequently only the input data set was recovered), resulting in high-specificity yet low-sensitivity predictions. However, the speed of ERPIN was comparable with that of the sequence-based methods.

The results for RaveNnA were also good, with the algorithm ranking third after Infernal and RSEARCH in terms of accuracy. The accuracy of RaveNnA when compared with the other profile HMM methods, HMMer and SAM, was excellent. The speed of RaveNnA was about the same magnitude as SAM, which is in good agreement with theory. However, RaveNnA requires a significant initialization time (~25 min, see Supplemental Fig. 5) from the overhead for calibrating the HMM to determining an appropriate threshold; therefore, it is only economical to use RaveNnA on larger databases.

The speed of RSmatch was nearly an order of magnitude greater than that of the structure-enhanced methods Infernal, RaveNnA, and RSEARCH; however, the accuracy was much lower.

Five versus twenty input sequences

Overall, the results were rather constant between using 5 or 20 input sequences. RSEARCH and Infernal exchanged first place; Infernal explicitly uses covariation information and hence is likely to be more powerful with larger input data sets. The performance of WU-BLAST dropped relative to the other programs.

The CM and profile HMM methods benefit from models derived from more sequences, resulting in improved specificity. The single-sequence methods, however, suffer from problems due to multiple testing, resulting in improved sensitivity at a cost to specificity.

65% scoring scheme

This study showed that the sequence-based methods perform rather similarly when using comparable parameter settings (Fig. 3; the results labeled "65%" in Supplemental Table 3; Supplemental Fig. 6). The nonheuristic method, SSEARCH, outranked the other methods in all cases. This was followed by FASTA. The two incarnations of BLAST performed almost identically; however, WU-BLAST was significantly faster than NCBI-BLAST.

RNA-centric scoring schemes

Each of the RNA-centric scoring schemes mentioned in the Methods section was given a trial (Fig. 4; Supplemental Table 4; Supplemental Fig. 7). The scoring schemes we tested are the PUPY matrix that ships with WU-BLAST, the "-U" option for FASTA, and the single-sequence components of the score matrices used by RSEARCH (Klein and Eddy 2003) and FoldAlign (Havgaard et al. 2005). These results were generally disappointing: None of the methods showed any improvement over less-specific schemes when the RNA-centric scores were used. In the case of the FoldAlign and RSEARCH score matrices, this is justified as these matrices were built specifically for structural methods rather than the sequence-based methods we have used here. We also tested a transition/transversion scoring scheme optimized for 65% sequence identity (States et al. 1991) (data not shown); the results of this test were also disappointing. This indicates that a great deal more work is required before such scoring schemes can be used for practical RNA homology search.

Application of phylogenetic sequence reconstruction to homology search

In general, the inclusion of ancestral information did not increase the performance of the more advanced methods. However, a number of methods did benefit from this approach. First, a significant improvement in the performance of ERPIN was observed for both the ASR and PSR approaches; the median sensitivity improved by a factor of 17 when PSR sequences were included in the search. One difficulty with this approach is the use of a prior on the length of the branch leading to the recon-

structed sequence. We found the best results with very long prior branch lengths (≈20 expected substitutions per site). In addition to improving ERPIN, we found that the inclusion of ASRs improved the sensitivities of the sequence-based methods. This improvement, unfortunately, comes at a cost to specificity, but in the cases of FASTA and particularly SSEARCH, an overall increase in accuracy was observed (Fig. 4). Our approach for including phylogenetic information did not improve the accuracy of the profile HMM or Infernal searches where tree-weighting schemes and Dirichlet priors are used. Since these search algorithms already incorporate information similar to that obtained from the ancestral sequences, a failure to see an increase in these methods is perhaps not surprising. Our results do suggest that future work may benefit from incorporating directly phylogenetic information along with the use of improved models of RNA sequence evolution. However, further work will be required to determine whether improvements in accuracy are due to additional information from the phylogeny or to noise injection (Krogh et al. 1994).

Genome scan

To test a representative set of algorithms in a realistic usage scenario, we scanned a 40-Mb section of the human genome. This test had the additional advantage of providing a more accurate estimate of algorithm specificity. The results of the genome scan were in general agreement with the earlier results. This was a much harder test than those presented earlier. This was compounded by the fact that the genome annotation we rely on may not be completely accurate. However, HMMer performed surprisingly well on this test compared with both Infernal and SAM. Infernal had the highest median MCC, yet had a slightly lower specificity than HMMer. This behavior may be due to unannotated homologs residing within the genomic region, which would cause false false-positives for the more sensitive methods such as Infernal and SAM. Of the single-sequence methods, both WU-BLAST and FASTA performed well. WU-BLAST had a higher sensitivity but a lower specificity (Fig. 4; Supplemental Fig. 8).

Discussion

The most popular homology search methods did not necessarily perform the best in our study. These programs are optimized for rapid database searches with few false positives (high specificity), which is not always what the user requires. As a consequence many estimates of the amount of conserved DNA (Hillier et al. 2004) and number of conserved ncRNAs (Washietl et al. 2005; Pedersen et al. 2006) and nonconserved ncRNAs (Pang et al. 2006) are based on suboptimal homology search tools and hence likely to be inaccurate.

One of the most important issues for homology searching is to develop scoring schemes that discriminate signal from noise. It is clear that the popular single sequence methods do not do this, although some modest improvements may be possible. The scorings implemented in the RNA-specific probabilistic methods Infernal, RSEARCH, and RaveNnA, however, do a good job of discriminating signal from noise. The Infernal CM method was surprisingly robust to the predicted (and therefore potentially inaccurate) input alignments and secondary structures. A comparison of Infernal with predicted and reference input data shows that the only time the predictions caused a drop in performance was when the input sequences were highly similar and the secondary structure prediction was poor. This meant that there was limited covariation information from the alignment for either the secondary structure prediction tool (RNAalifold) or the covariance model to use (see Supplemental Fig. 10).

As researchers are usually likely to favor speed over accuracy, it is necessary that FASTA and BLAST have accurate scoring schemes available that such researchers can utilize. For this, RNA-optimized PAM (Dayhoff et al. 1978) and/or BLOSUM (Henikoff and Henikoff 1992) style score matrices are needed. There are sufficient data for computing these matrices from freely available sources such as the Rfam (Griffiths-Jones et al. 2003) and the rRNA databases (Cannone et al. 2002; Wuyts et al. 2002) and the statistical methods for estimating these are well established. Additionally, given that base-pair stacking is important for RNA structure, this signal may prove useful for RNA homology search and could be exploited by incorporating a dinucleotide scoring scheme into the alignment procedure (Lunter and Hein 2004).

There are few heuristics at present for rapid profile HMM and CM-based homology searches. One could, for example, apply the BLAST concept of a seed match to profile HMMs and CMs. A database could be rapidly scanned for short, ungapped high-scoring matches to the model, which could then be extended using the full profile HMM architecture, this should result in significant gains in speed at moderate costs to sensitivity.

The specificity of homology search algorithms is an important issue when scanning large amounts of (genomic) data. If the specificity is not extremely high when scanning large databases the relatively small number of true homologs may be lost in a flood of false positives. The relatively small amounts of shuffled data we were forced to use in this study (because of the glacial speeds of some of the test algorithms) meant that we did not get an accurate measure of this value. The genome scan test we ran was meant to alleviate this problem; however, it is likely that unannotated homologs also affected the determination of algorithm specificity.

The use of phylogenetic information to enhance homology search of ncRNAs on the surface seems a bit dissatisfying. Our results do clearly indicate that there is valuable information in the phylogeny that should not be ignored as a number of the methods did benefit from our simple approach. We believe this suggests that future method development will benefit from considering the phylogeny when multiple sequences are available. For example, the use of mutational maps along the phylogeny (Nielsen 2002) could be used to create a stochastic profile for profile HMM methods (Durbin et al. 1998). Alternatively, when scanning newly sequenced genomes we often know the phylogenetic relationships between the search sequences and the query genome and may also have information on how divergent they are. Using this information, one could use the PSR approach described here to add information to the search sequences without relying on arbitrary priors about the process of evolution. Our results do hold promise for those researchers who want the set of putative homologs to include all true homologs at the cost of including a larger number of false positives.

The RSmatch algorithm relies on MFE structure predictions on a single sequence, which are known to be frequently inaccurate (Gardner and Giegerich 2004). If the structure prediction phase for both the database and input sequences were based on comparative predictions, such as RNAalifold (Hofacker et al. 2002), the accuracy of this approach is likely to improve. In addition, RSmatch could be used to cluster genome-wide structure-based ncRNA predictions (Washietl et al. 2005; Pedersen et al. 2006).

Practical recommendations

On the basis of our assessment of the currently available homology search tools, we recommend a scheme for one or more input sequences that uses iterative rounds of the rapid sequence-based methods, such as WU-BLAST, FASTA, or SSEARCH, with sensible scoring schemes and a high threshold to build a training data set. These results can then be used to train CM models for Infernal searches to obtain more divergent sequences from within the lower-scoring sequence-based matches or, if possible, the original database. Profile HMMs, particularly RaveNnA, could be used instead when CMs are not practical, for example, when the sequence length is greater than 200 nucleotides or the database is large.

Throughout this work, we have focused almost entirely on method accuracy; however, of frequent concern is the computation time for a search. For example, on the basis of our timings with tRNA queries, Infernal would take ~96 d to search the human genome on a single processor, RaveNnA would take 40 h, HMMer would take 9 h, SSEARCH would take 4 h, and WU-BLAST (W7) just 4 min. Given the ready access many groups have to computing clusters, it is reasonable to expect the more accurate methods to become more popular in the future.

Many of the currently available tools for ncRNA homology search tools are not yet performing as well as one would hope based on the results we have presented. Improvements in terms of accuracy and speed are needed. This is extremely important given the explosion of interest in ncRNAs generated by recently discovered ncRNAs such as miRNAs. Additionally, a current theory suggests that much of the apparent organism complexity not accounted for by corresponding expansions in the proteome can be attributed to regulation from the ncRNAome (Mattick 2001).

This study has implications for evolutionary studies that rely on programs tuned for high-similarity sequences. First, since the most popular programs are biased toward identifying only highly conserved homologs, the diversity of particular ncRNAs will be underestimated perhaps more severely than previously thought. Second, and particularly irksome, is that many of the most interesting homologs will be those that are or have experienced strong positive directional or diversifying selection, causing them to have diverged beyond the detection limits of the search algorithms, and will fail to be identified. By establishing how the currently available methods perform, we can gather more high-quality databases that will allow further development of our understanding as to how different families of ncRNAs evolve. With new models in hand, we can improve the search algorithms increasing the discovery of interesting homologs, otherwise unidentified, and gain better estimates of the diversity of ncRNAs across the spectrum of life.

Methods

In the following section, we outline the data sets we used for this study and the approaches we used to compare sequence-based, profile HMM, and structure-enhanced methods.

Data sets

To test homology search tools, we have obtained hand-curated databases of 602 5S rRNAs, 1114 tRNAs, and 235 U5 spliceosomal RNAs. Sequences in the databases have mean lengths of 117, 73, and 119 nucleotides, respectively (Zwieb 1997; Sprinzl et al. 1998; Szymanski et al. 2002; Griffiths-Jones et al. 2003). Nonhomologs were generated by shuffling sequences from each database to generate a new database that was 10 times larger. The shuffling process preserved dinucleotide frequencies to avoid creating artificially dissimilar sequences (Workman and Krogh 1999). Sets of 5 and 20 search sequences were sampled from the databases. These were used to scan the original and a shuffled database. A total of 583 5-sequence sets and 360 20-sequence sets were generated.

Performance measures

Three measures of performance were used to evaluate each algorithm: sensitivity, specificity, and MCC. Sensitivity measures the fraction of the positive control data set that is recovered by an algorithm and is calculated from the number of matches to the unshuffled database for both 5 and 20 sequence sets as inputs. Specificity measures the fraction of the randomized sequences that were correctly rejected when scanning the shuffled databases with the input sequence sets of 5 and 20. The third measure, MCC, combines both sensitivity and specificity to measure the overall discriminative power of each algorithm.

Thresholds

Thresholds are used to provide a cutoff for determining whether a query sequence matches the search sequence(s). Score thresholds for each algorithm are optimized based on scans of the curated and shuffled databases with a small group of query sequences that uniformly covers the different RNA families and identity ranges. Example distributions and ROC plots are illustrated in Supplemental Figures 3 and 4. Raw scores rather than e-values are used here as there are a diverse number of methods implemented for computing e-values and these are not computed at all by some methods. In this study, we are more concerned with the scores used by specific programs than the accuracy of the different e-value computations.

65% scoring scheme

To compare the sequence-based methods on an equal footing, we have included a comparison of these using parameters optimized for data sets with 65%–100% homology (States et al. 1991) (match = +5, mismatch = −4, gapopen = 10, gapextension = 10). Where a seed was required, this was made as similar as possible, W = 7 for BLAST and ktup = 6 for FASTA.

RNA-centric scoring schemes

Several of the sequence-based methods have associated scoring schemes that are designed for the unique problem of RNA homology search. Generally, these distinguish between transitions (A ↔ G and C ↔ U), which are relatively frequent during RNA evolution, and transversions (the remaining mismatch types), which are relatively infrequent. WU-BLAST has a PUPY (purine–pyrimidine) score matrix (match = +4, transition = +2, transversion = −8). By default, FASTA and SSEARCH score a +5 for matches and −4 for mismatches, yet these tools have a "−U" scoring option that tolerates G · U wobble base pairs by scoring G/A and U/C mismatches as one less than a G/G match in a strand-specific manner. In addition, RSEARCH's RIBOSUM (Klein and Eddy 2003) and the more recent FoldAlign (Havgaard et al. 2005) score matrices use parameters estimated directly from the loop regions of large curated ncRNA alignments. We have tested these as an alternative scoring scheme for FASTA homology searches.

Genome scan

A ncRNA-enriched region of the human genome was selected for further testing of representative homology search programs from each category. We identified a 40-Mb region on chromosome 12 (coordinates 90,000,000–130,000,000; genome assembly NCBI35) that contains 5 5S rRNAs, 10 tRNAs (and 26 pseudo-tRNAs predicted by tRNAscan-SE), and 1 U5 spliceosomal RNA. We used input data sets containing ten sequences, each with a sequence identity to the associated target RNA in one of the following ranges 40%–60%, 50%–70%, 60%–80%, and 70%–85%. The pairwise sequence identities within the data sets are between 60% and 90%.

Timing

For the timing studies, two databases of 166 Mb and 332Mb, respectively, were used. Both databases contain 1114 tRNA sequences; the smaller database has one shuffled version of each of these, whereas the larger database has three. A single tRNA subset was used as a query for the timing study. The times for scanning the 2 databases are computed on (or calibrated to) a Sun Sparc v9 and 900 MHz CPU for each algorithm. From these values, the algorithm speed (nts/s) and initialization times are computed.

Phylogenetic information

The phylogenetic relationships among the search sequences include information on the branching order and their evolutionary divergence. It has been previously suggested that ancestral reconstruction of the sequences can be used to aid homology searches by supplementing the search set with reconstructed sequences (Collins et al. 2003; McCormack 2003; Qian and Goldstein 2003; Cai et al. 2004). To test whether this type of phylogenetic information will aid in identifying homologous ncRNAs, we used an empirical Bayesian approach (Huelsenbeck and Bollback 2001) to stochastically sample ancestral sequences (ASR) (details of the sampling can be found in the Supplemental Methods). The Bayesian approach has been proven (in the case of proteins) to be the most accurate method (Hall 2006). The phylogenetic tree and model parameters were estimated using MrBayes v3 (Huelsenbeck and Ronquist 2001). To accommodate the nonindependence among sites arising from secondary structure, an RNA doublet model was used to model substitutions in stem regions (Schöniger and von Haeseler 1994; Huelsenbeck and Ronquist 2001), while loop regions were modeled using the method of Hasegawa et al. (1985). In addition to reconstructing ancestral sequences at the internal nodes of the phylogeny, a novel simple method was employed to sample ancestral sequences from unobserved lineages radiating from the internal nodes of the phylogeny using a Bayesian posterior predictive (PSR) approach (Bollback 2005) (see Supplemental Methods).

Alignment and structure prediction

We used automatic structure prediction and alignment methods that previous studies have identified as being accurate for ncRNA analyses (Gardner and Giegerich 2004; Gardner et al. 2005). The alignments are computed using ProAlign (Löytynoja and Milinkovitch 2003) and consensus structures are computed from these alignments using RNAalifold (Hofacker et al. 2002).

Performance measures

Sensitivity and *specificity* are common measures for determining the accuracy of homology search methods.

$$Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP}$$

where *TP* is the number of "true positives," *TN* is the number of "true negatives," *FN* is the number of "false negatives," and *FP* is the number of "false positives." *Sensitivity* measures the fraction of the positive control data set that is recovered by the program in question; the *specificity* measures what fraction of the randomized sequences that were correctly rejected.

A measure combining both specificity and sensitivity is useful for ranking programs. In previous studies, the

$$error\ rate = \frac{FP + FN}{TP + TN + FP + FN}$$

(Klein and Eddy 2003) has been used; we, however, favor the more discriminative Matthews' correlation coefficient (MCC) as defined below:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The MCC ranges from -1 for extremely inaccurate ($TP = TN = 0$) to 1 for very accurate predictions ($FP = FN = 0$).

In general, we measure *TP* as the number of unique sequences in the hand-curated databases that were accepted by the algorithm in question using each of the 5 or 20 input sequences. For the genome scan, TP is instead measured as the number of nucleotides that are in both a known and a predicted RNA sequence. From this it follows how TN, FP, and FN are computed in each case. To make a distinction between the regular performance measures defined here and the ones used for the genome scan, we call the latter *nSensitivity*, *nSpecificity*, and *nMCC*.

To ease the comparison of the different measures, we have computed the rank of each program with representative parameter settings against the other programs using MCC values for each subset of query sequences. The rank distributions are plotted in Figures 2–4.

Acknowledgments

We thank Sam Griffiths-Jones, David Ardell, Anders Krogh, Rasmus Nielsen, Zasha Weinberg, and Jeppe Vinther for useful discussions. We also thank the homology-search algorithm developers Torbjørn Rognes, Bin Tian, William R. Pearson, Robert J. Klein, Zasha Weinberg, Stephen Altschul, Daniel Gautheret, and Sean Eddy for taking the time to make useful comments on an early draft of this manuscript. Any remaining flaws are solely our responsibility. The high-performance computer clusters at UPPMAX and the University of Copenhagen Bioinformatics Centre were used to compute many of the results presented here. P.P.G. is supported by a Carlsberg Foundation Grant (21-00-0680). J.P.B. was supported by a grant from the Danish FNU.

References

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

Bollback, J.P. 2005. Posterior mapping and predictive distributions. In *Statistical Methods in Molecular Evolution* (ed. R. Nielsen), pp. 189–203. Springer Verlag, New York.

Brenner, S., Chothia, C., and Hubbard, T. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* **95:** 6073–6078.

Cai, W., Pei, J., and Grishin, N.V. 2004. Reconstruction of ancestral protein sequences and its applications. *BMC Evol. Biol.* **4:** 33.

Cannone, J., Subramanian, S., Schnare, M., Collett, J., D'Souza, L., Du, Y., Feng, B., Lin, N., Madabusi, L., Muller, K., et al. 2002. The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3:** 2.

Chao, K.M., Pearson, W.R., and Miller, W. 1992. Aligning two sequences within a specified diagonal band. *Comput. Appl. Biosci.* **8:** 481–487.
Collins, L.J., Poole, A.M., and Penny, D. 2003. Using ancestral sequences to uncover potential gene homologues. *Appl. Bioinformatics* **2:** 85–95.
Dayhoff, M., Schwartz, R., and Orcutt, B. 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345–352. National Biomedical Research Foundation, Washington, D.C.
Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of protein and nucleic acids*. Cambridge University Press, Cambridge, UK.
Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.
Eddy, S.R. 2002. A memory efficient dynamic programming algorithm for optimal structural alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **3:** 18.
Eddy, S.R. and Durbin, R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22:** 2079–2088.
Gardner, P.P. and Giegerich, R. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* **5:** 140.
Gardner, P.P., Wilm, A., and Washietl, S. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.* **33:** 2433–2439.
Gautheret, D. and Lambert, A. 2001. Direct RNA definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.* **313:** 1003–1011.
Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. 2003. Rfam: An RNA family database. *Nucleic Acids Res.* **31:** 439–441.
Hall, B.G. 2006. Simple and accurate estimation of ancestral protein sequences. *Proc. Natl. Acad. Sci.* **103:** 5431–5436.
Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human–ape splitting by molecular clock of mitochondrial DNA. *J. Mol. Evol.* **21:** 160–174.
Havgaard, J.H., Lyngsø, R., Stormo, G.D., and Gorodkin, J. 2005. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics* **21:** 1815–1824.
Haussler, D., Krogh, A., Mian, I.S., and Sjölander, K. 1993. Protein modeling using hidden Markov models: Analysis of globins. In *Proceedings of the Hawaii International Conference on System Sciences*, pp. 792–802. IEEE Computer Society Press, Los Alimitos, CA.
Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.
Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.
Hofacker, I.L., Fontana, W., Bonhoeffer, S., and Stadler, P.F. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125:** 167–188.
Hofacker, I., Fekete, M., and Stadler, P. 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319:** 1059–1066.
Huelsenbeck, J.P. and Bollback, J.P. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Syst. Biol.* **50:** 351–366.
Huelsenbeck, J.P. and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17:** 754–755.
Hughey, R. and Krogh, A. 1996. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *Comput. Appl. Biosci.* **12:** 95–107.
Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14:** 846–856.
Klein, R.J. and Eddy, S.R. 2003. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics* **4:** 44.
Krogh, A., Brown, M., Mian, I.S., Sjölander, K., and Haussler, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235:** 1501–1531.
Lindahl, E. and Elofsson, A. 2000. Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295:** 613–625.
Liu, J., Wang, J.T., Hu, J., and Tian, B. 2005. A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics* **6:** 89.
Löytynoja, A. and Milinkovitch, M.C. 2003. A hidden Markov model for progressive multiple alignment. *Bioinformatics* **19:** 1505–1513.
Lunter, G. and Hein, J. 2004. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* **20:** I216–I223.
Madera, M. and Gough, J. 2002. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.* **30:** 4321–4328.
Mattick, J. 2001. Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Rep.* **2:** 986–991.
McCormack, T.J., 2003. Comparison of K^+-channel genes within the genomes of *Anopheles gambiae* and *Drosophila melanogaster*. *Genome Biol.* **4:** R58.
Nielsen, R. 2002. Mapping mutations on phylogenies. *Syst. Biol.* **51:** 729–732.
Pang, K.C., Frith, M.C., and Mattick, J.S. 2006. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends Genet.* **22:** 1–5.
Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284:** 1201–1210.
Pearson, W.R. 1991. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11:** 635–650.
Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.
Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., and Haussler, D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2:** 251–262.
Qian, B. and Goldstein, R.A. 2003. Detecting distant homologs using phylogenetic tree-based hmms. *Proteins* **52:** 446–453.
Saebø, P.E., Andersen, S.M., Myrseth, J., Laerdahl, J.K., and Rognes, T. 2005. PARALIGN: Rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.* **33:** 535–539.
Schöniger, M. and von Haeseler, A. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.* **3:** 240–247.
Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., and Haussler, D. 1996. Dirich-let mixtures: A method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.* **12:** 327–345.
Smith, T. and Waterman, M. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.
Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. 1998. Compilation of trna sequences and sequences of trna genes. *Nucleic Acids Res.* **26:** 148–153.
States, D.J., Gish, W., and Altschul, S.F. 1991. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods Enzymol.* **3:** 66–70.
Szymanski, M., Barciszewska, M.Z., Erdmann, V.A., and Barciszewski, J. 2002. 5S Ribosomal RNA Database. *Nucleic Acids Res.* **30:** 176–178.
Washietl, S., Hofacker, I.L., Lukasser, M., Hüttenhofer, A., and Stadler, P.F. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* **23:** 1383–1390.
Weinberg, Z. and Ruzzo, W.L. 2006. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics* **22:** 445–452.
Workman, C. and Krogh, A. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.* **27:** 4816–4822.
Wuyts, J., Van de Peer, Y., Winkelmans, T., and De Wachter, R. 2002. The European database on small subunit ribosomal RNA. *Nucleic Acids Res.* **30:** 183–185.
Zwieb, C. 1997. The uRNA database. *Nucleic Acids Res.* **25:** 102–103.

Received August 23, 2006; accepted in revised form October 19, 2006.