Homology Search with Fragmented Nucleic Acid Sequence Patterns

Axel Mosig¹⁸, Julian J.-L. Chen²³, and Peter F. Stadler⁴⁵⁶⁷

¹ Department of Combinatorics and Geometry (DCG), CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences (SIBS) Campus, Shanghai, China

axel@picb.ac.cn

² School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA
jlchen@asu.edu

³ Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287, USA

⁴ Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany;

⁵ RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie — IZI
Deutscher Platz 5e, D-04103 Leipzig, Germany;

⁶ Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria;

⁷ Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
studla@bioinf.uni-leipzig.de

⁸ Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany

Abstract. The comprehensive annotation of non-coding RNAs in newly sequenced genomes is still a largely unsolved problem because many functional RNAs exhibit not only poorly conserved sequences but also large variability in structure. In many cases, such as Y RNAs, vault RNAs, or telomerase RNAs, sequences differ by large insertions or deletions and have only a few small sequence patterns in common.

Here we present **fragrep2**, a purely sequence-based approach to detect such patterns in complete genomes. A **fragrep2** pattern consists of an ordered list of position-specific weight matrices (PWMs) describing short, approximately conserved sequence elements, that are separated by intervals of non-conserved regions of bounded length. The program uses a fractional programming approach to align the PWMs to genomic DNA in order to allow for a bounded number of insertions and deletions in the patterns. These patterns are then combined to significant combinations of PWMs. At this step, a subset of PWMs may be deleted, i.e., have no match in the current region of the genome. The program furthermore estimates *p*- and *E*-values for the matches.

We apply **fragrep2** to homology searches for RNase MRP, unveiling two previously unidentified matches as well as reproducing the results of two previous surveys. Furthermore, we complement the picture of vertebrate vault RNAs, a class of ncRNAs that has not received much attention so far.

1 Introduction

With the rapidly increasing number of completely sequenced genomes their annotation becomes an important task in bioinformatics. In contrast to protein coding genes, fast and reliable annotation of ncRNAs is still in its infancy with few exceptions such as tRNA annotation using **tRNAscanSE** [1]. The reason is that most families of ncRNAs are short and rather variable at sequence level. Often secondary structure is well conserved; in this case structure-based tools such as **infernal** [2] can be used. These approaches are too slow to be used on their own, however. Consequently, the ENSEMBL ncRNA pipeline uses a non-restrictive **blast** search to filter likely regions that are then searched with **infernal**. Similarly, in [3] a sequence-based filtering step using an HMM-based heuristic is proposed.

For many ncRNA families, dramatic variations in sequence length pose an additional problem. In conjunction with rapid substitutions it becomes surprisingly hard to find homologs even of fairly long ncRNAs. For example, telomerase RNAs (with a length of more than 300nt) cannot be found by means of **blast** in the genomes of teleost fishes, even though more than 20 sequences representing all other major vertebrate clades are known [4]. On the other hand, most ncRNAs exhibit a few short well conserved sequence motifs. The **fragrep** program [5] has been designed to exploit this fact: in its original version it searches for (almost) exact sequence patterns separated by non-conserved sequence intervals which are specified only in terms of upper and lower bounds on their length.

In practical applications we found, however, that almost exact gap-less patterns are too restrictive to search in evolutionarily distant genomes. A relaxed version, using a heuristic to match PWMs to genomic DNA, was indeed employed successfully for discovering telomerase RNAs in teleosts [6]. This work also showed that entire conserved blocks can be lost (or mutated beyond recognition) in certain lineage, so that it is important to allow for the “deletion” of entire sub-patterns. In this contribution we describe a novel version of **fragrep** which (1) uses an exact fractional programming algorithm for matching PWMs to genomic DNA with insertions and deletions and (2) accounts for deleted sub-patterns as well as missing data at the ends of contigs in incomplete sequence assemblies.

2 Theory

A **fragrep** pattern, Fig. 1, is specified as a vector of triplets $([u_i, v_i], \mathbf{W}^i, d_i | i = 1, \dots, N)$ where $[u_i, v_i]$ denoted the upper and lower bound of the number of nucleotides between the patterns $\mathbf{W}^{i-1}$ and $\mathbf{W}^i$, which are specified as PWMs, and d_i is the maximal number of in/dels that are tolerated in a match of $\mathbf{W}^i$ to the genomic sequence. The sequence interval before the first pattern need not be specified, hence we set $[u_1, v_1] = [0, 0]$.

The problem of finding a **fragrep** pattern in a genomic sequence x consists of two parts: (1) finding matches of the individual patterns $\mathbf{W}^i$ in the genome, and (2) combining such matches that satisfy the distance constraints.

Fig. 1. Example of a **fragrep** pattern for vault RNAs derived from 40 tetrapod vault RNA sequences retrieved from the Rfam database, ENSEMBL 43 annotation, and **blast** searches of additional vertebrate genomes.

2.1 Position-Specific Weight Matrices with InDels

Let $\mathbf{W} = (W_{k\alpha})$ be a position specific weight matrix (PWM) representing a pattern of length n over the alphabet $\mathcal{A}$, i.e., $k \in [1, n]$ and $\alpha \in \mathcal{A}$, normalized such that $\sum_{\alpha} W_{k\alpha} = 1$. Following the slight abuse of notation from literature on modelling transcription factor binding sites, we refer to these normalized frequency count matrices as PWMs. One common way [7] of scoring the match of $\mathbf{W}$ with a sequence $(x_1, \dots, x_n)$, $x_i \in \mathcal{A}$ is to compute position-wise scores of the form

$$s_i = H_i W_{i, x_i} \quad s_i^+ = H_i \max_{\alpha \in \mathcal{A}} W_{i, \alpha} \quad s_i^- = H_i \min_{\alpha \in \mathcal{A}} W_{i, \alpha} \quad (1)$$

where $H_i = \sum_{\alpha \in \mathcal{A}} W_{i\alpha} \ln W_{i\alpha}$ is the information content of column i . These scores are then combined to the so-called *MATCH* score [7]:

$$S = \left(\sum_{i=1}^n s_i - \sum_{i=1}^n s_i^- \right) / \left(\sum_{i=1}^n s_i^+ - \sum_{i=1}^n s_i^- \right). \quad (2)$$

Here we generalize this approach to incorporate insertions and deletions. We observe that this problem is inherently asymmetric, i.e., the effect of deleting a letter in the sequence x is not the same as that of deleting a column of the PWM. For a given “alignment path” $\mathbb{A}$, i.e., a sequence of matches (i, i') , insertions $(-, i)$ and deletions $(i, -)$, we modify the initial scores by setting $s_{i,j} = H_i W_{i, x_j}$ for the match of the i -th column of the PWM with the j -th letter in x . Introducing gap penalties Δ_s for the sequence and Δ_{pwm} for the PWM, respectively, we evaluate a given alignment $\mathbb{A}$ path by means of

$$S(\mathbb{A}) = \left(\sum_{\text{matches}(i,j)} s_{i,j} - \sum_{j:(-,j)} \Delta_s - \sum_{i:(i,-)} \Delta_{pwm} H_i - \sum_{i:(i,*)} s_i^- \right) / \left(\sum_{i:(i,*)} s_i^+ - \sum_{i:(i,*)} s_i^- \right). \quad (3)$$

If there are no indels, $S(\mathbb{A})$ coincides with the usual *MATCH* score. Furthermore, $S(\mathbb{A}) \leq 1$. The score can, however, become negative if there are too many insertions or deletions.

We follow the general framework of *fractional programming* [8, 9]: since optimizing a fractional target function is often difficult, one turns the optimization problem into the corresponding decision problem. As it turns out in many cases, this decision can be solved by a non-fractional optimization problem. The optimal value of the original fractional optimization problem can then be found using nesting intervals. In our case, we ask whether there is an alignment path $\mathbb{A}$ such that $S(\mathbb{A}) \geq \vartheta$ with a given threshold $\vartheta \leq 1$. This decision problem is satisfied if and only if there is an alignment path $\mathbb{A}$ such that

$$\sum_{\text{matches}(i,j)} \underbrace{s_{i,j} - s_i^- - \vartheta(s_i^+ - s_i^-)}_{\sigma_{ij}(\vartheta)} - \sum_{\text{insertion}(-,j)} \Delta_s - \sum_{\text{deletion}(i,-)} \Delta_{pwm} H_i \geq 0 \quad (4)$$

The decision problem (4) is closely related to normalized editing problems, which can be solved in an analogous way, see [10, 11] for details. In order to decide whether (4) has a non-negative solution, it suffices to solve an asymmetric variant of the standard sequence alignment problem with match scores that depend explicitly on the parameter ϑ . For fixed ϑ this can be done by means of the following asymmetric variant of the Needleman-Wunsch algorithm [12]:

$$S_{ij} = \begin{cases} S_{i-1,j-1} + \sigma_{ij}(\vartheta) \\ S_{i-1,j} - \Delta_{pwm} H_i \\ S_{i,j-1} - \Delta_s \end{cases} \quad (5)$$

with initial condition $S_{0,0} = 0$ and the convention that alternatives with negative indices are ignored. Following the logic of earlier versions of **fragrep** [5] we optionally restrict the number of insertions and deletions. This leads to the following simple generalization of algorithm (5). Let d be the number of deleted PWM columns in the alignment path before i, j and denote by $S_{i,j;d}$ the optimal score with of an alignment up to positions i and j , resp., with exactly d deletions in the PWM. One easily verifies that these quantities satisfy the recursion

$$S_{i,j;d} = \begin{cases} S_{i-1,j-1;d} + \sigma_{ij}(\vartheta) \\ S_{i-1,j;d-1} - \Delta_{pwm} H_i \\ S_{i,j-1;d} - \Delta_s \quad \text{if } j - (i + d) \leq d_{\text{threshold}} \end{cases} \quad (6)$$

with initial condition $S_{0,0;0} = 0$. Note that $j - (i + d)$ equals the number of deletions from the sequence within the alignment path.

Denote the optimal score of this dynamic programming step, with or without constraint on the maximum number d of insertions and deletions, by $S[\vartheta]$. Fractional programming [8, 9] is based on the observation that $S[\vartheta] \geq 0$ if and only if the fractional score can exceed ϑ . This implies that the optimal value of ϑ can be obtained by means of the following iteration:

$$\vartheta \leftarrow \vartheta + \text{sgn} S[\vartheta] \times \Delta\vartheta \quad \Delta\vartheta \leftarrow \Delta\vartheta/2 \quad (7)$$

with initial condition $\vartheta = 1$ and $\Delta\vartheta = 1 - \vartheta^*$ where ϑ^* is the *a priori* threshold below which a match of $\mathbf{W}$ with x is deemed insignificant.

2.2 Significance Statistics

Once the optimal value of ϑ is determined, we can retrieve a corresponding optimal alignment path by means of backtracking in the most recently computed S -matrix. We either use ϑ as the weight of the match, or we may employ a p or E value statistic for this purpose.

The probability p that the PWM $\mathbf{W}$ of length n matches a random sequence (α_i) of length n as least as well as the actual sequence $x = (x_1, \dots, x_n)$ can be computed as

$$p(\mathbf{W}, x) = \text{Prob} \left[u := \sum_{i=1}^n H_i W_{i, \alpha_i} > \sum_{i=1}^n H_i W_{i, x_i} \right] \quad (8)$$

where the letters α_i are drawn i.i.d. from the background distribution corresponding to the genomic frequencies π_α . By the central limit theorem the random variable u will follow at least approximately a Gaussian distribution, whose mean μ and variance σ^2 is given by

$$\mu = \sum_{i=1}^n H_i \sum_{\alpha} W_{i, \alpha} \pi_{\alpha} \quad \sigma^2 = \mu^2 - \sum_{i=1}^n H_i^2 \sum_{\alpha} W_{i, \alpha}^2 \pi_{\alpha} \quad (9)$$

and hence depends only on the PWM $\mathbf{W}$ and on the background nucleotide frequencies π_α .

In the case of indels we match a sequence of $n + I$ with a PWM pattern of length $n - D$, where I and D are the numbers of insertions and deletions respectively. Thus we can estimate the probability of match with indels at least roughly as $p' \approx \binom{n+I}{n-D} p(\mathbf{W}, x)$. As in the first version of **fragrep** [5], we account for the variable length intervening sequences by considering all combinations of the length of intervening sequence pieces, i.e.,

$$p \approx p'(\mathbf{W}^1, x^1) \prod_{k=2}^N p'(\mathbf{W}^k, x^k) (v_k - u_k + 1) \quad (10)$$

While by no means an exact p -value, this approximation is good enough to rank-order the hits of a database search, to allow a comparison of results obtained with different patterns, and to weed out obviously insignificant results.

2.3 Optimal Combinations of PWM Patterns

As indicated above, we are interested in combinations of PWM occurrences that satisfy certain distance constraints. To formalize this further, we assume that for each PWM $\mathbf{W}_i$, a threshold θ_i as well as a maximum number of deletions d_i

is specified. Given these constraints, we now consider all those pairs (i, ν) where $\mathbf{W}_i$ achieves a match score of at least t_i with at most d_i deletions at position ν in the genomic sequence to be searched; we will refer to index pairs (i, ν) satisfying these constraints as *occurrences*.

Denote by (i, ν) an occurrence of PWM pattern $\mathbf{W}_i$ at genomic sequence position ν with weight $w_{(i, \nu)}$. We say that (i', ν') is a *predecessor* of (i, ν) , in symbols $(i', \nu') \prec (i, \nu)$, if the following condition is satisfied:

$$(i', \nu') \prec (i, \nu) \Leftrightarrow \sum_{j=i'}^i u_j \leq \nu - \nu' \leq \sum_{j=i'}^i (v_j + n_j + d_j) \quad (11)$$

where n_i is the length of the pattern $\mathbf{W}_i$.

Rather than requiring all PWMs in the search pattern to be matched, we allow a fixed (and typically small) number of δ many patterns to be deleted in a matching sequence. In other words, we are interested in finding chains of length $k - \delta$ in the above (partial) order, k denoting the number of PWMs in our search pattern.

Now let $D_{(i, \nu)}$ be the minimum number of deletions associated with a path ending in (i, ν) . It satisfies the recursion

$$D_{(i, \nu)} = \min \begin{cases} \min_{(i', \nu') \prec (i, \nu)} [D_{(i', \nu')} + (i - i' - 1)] \\ i - 1 \\ 0 \text{ if } (0, 0) \prec (i, \nu) \end{cases} \quad (12)$$

Here $(0, 0)$ denotes the origin of the sequence. The last alternative allows us to ignore the deletion of partial pattern at the begin of a sequence. This is useful e.g. when dealing with unfinished genome assemblies or even unassembled shotgun reads. Matches for a given query can be derived in a straightforward manner from the entries in the table D : whenever we find $D_{(i, \nu)} \geq \delta - N - i$, we can identify a match by tracing back the corresponding chain of occurrences in the matrix D .

In practice, there can be a large number of very similar matches at essentially the same genomic location due to multiple occurrences of individual patterns $\mathbf{W}_i$ with low individual significance. To avoid redundant output, **fragrep2** has been implemented to report only the most significant match out of a bundle of essentially equivalent ones. Furthermore, **fragrep2** initiates dynamic programming of table D only at those locations where an occurrence of one of the $\delta + 1$ most significant patterns has been reported. As another measure of performance tuning, the partial order underlying table D is maintained in a dynamic data structure, so that no genomic location is ever searched more than once for any pattern $\mathbf{W}_i$.

3 Generating Search Patterns

The most convenient starting point for creating PWM-based search patterns are alignments of non-coding RNA families, such as those contained in databases

such as **Rfam** [13]. Conserved blocks with a length typically varying between three and 20 nucleotides are frequently observed in these alignments. These conserved blocks are easily identified and annotated by manual inspection. In order to facilitate the application of **fragrep2**, we developed a simple tool, **aln2pattern**, which reads an annotated sequence alignment and computes both the PWMs (from contiguous marked alignment columns) and the distance constraints between them (from the lengths of the sequences between marked blocks). The match score thresholds for the PWMs as well as column deletion bounds and distant constraints between the matrices are determined in such a way that they represent the *most restrictive search pattern that is capable of identifying all sequences in the given alignment*.

In a typical application scenario, these constraints will then be manually relaxed by the user, depending on how far the genome to be searched is evolutionarily away from the given training sequences in the multiple sequence alignment.

4 Applications

4.1 RNase MRP

RNase MRP is a ribonucleoprotein whose RNA component, MRP RNA, is enzymatically active. It is associated with at least two distinct roles in eukaryotes: (1) initiation of mitochondrial DNA replication and (2) cleavage of primary rRNA transcripts. RNase MRP is a distant relative of RNase P, a ribonucleoprotein involved in tRNA maturation. Two recent studies describe systematic searches for RNase MRP homologs in diverse eukaryotic genomes [14, 15]. In both studies combinations of manually composed sequence and secondary structure patterns were used for the search.

Fig. 2. Search patterns successfully used for finding RNase MRP: the pattern on the left identifies fungal species; the pattern on the right reveals unique and previously unidentified candidate matches in the amphioxus and *Saccoglossus kowalevskii*.

As a first test case for **fragrep2**, we used a pattern from known yeast RNase MRPs for whole-genome surveys in other yeast genomes. The pattern depicted

in Figure 2 (left) indeed retrieves salient matches in the genome sequences of *Ashbya gossypii*, *Candida glabrata*, *Debaryomyces hansenii*, *Kluyveromyces lactis*, and *Yarrowia lipolytica*. While most of these have been identified previously, the full advantage of the approach implemented in **fragrep2** becomes obvious through the MRP candidates we identified in the amphioxus and *Saccoglossus kowalevskii* using the pattern depicted in Figure 2 (right); **fragrep2** returns unique matches in both of these species, where RNase MRP has not been identified previously. This setting is particularly challenging, since all species with annotated MRP genes are phylogenetically quite distinct, and homology is consequently very spurious.

While the examples shown above reveal unique matches in a relatively direct way, the search procedure in more challenging genome-wide surveys eventually needs to be conducted in several stages: starting with low sensitivity, a long list of candidate matches is reduced by variants of the first pattern with lower fault tolerance (i.e., higher match thresholds for the individual PWMs). Final matches are then usually identified among one or few final candidates using either phylogenetic or secondary structure analysis using state-of-the-art tools.

4.2 Vault RNAs

Vault RNAs (vRNAs) are short (80-150nt) polymerase III transcripts that comprise about 5% of the vault complex. Vaults are large ribonucleoprotein particles that are ubiquitous in eukaryotic cells with characteristic barrel-like shape and a still poorly understood function in multi-drug resistance, see [16] for a review.

Vault RNAs exhibit little sequence conservation beyond their internal box A and box B internal promoter elements. So far, only a few examples have been studied experimentally. The human genome contains three expressed vRNAs located in a cluster on Chr.5 [17], while mouse and rat have only a single vRNA [18–20]. Outside mammals, only two vRNAs from the bullfrog *Rana catesbeiana* have been sequenced [18]. Due to their high sequence variability, ENSEMBL’s ncRNA annotation pipeline has identified vRNAs only in mammals and in *Xenopus* (due to its close relationship with the bullfrog). While additional mammalian vRNA sequences can easily be found by means of **blastn**, **blastn** searches outside the mammals have been unsuccessful.

Using **fragrep2** with the pattern in Fig. 1 we not only recover all vRNAs that are experimentally known or annotated in ENSEMBL (based on **infernal**), we also find plausible vRNA sequence in the genomes of chicken as well as zebrafish and tetraodon, which previously remained undetected⁹. In general, **fragrep2** returns a handful of candidates, among which the true homologs are easily recognized in a sequence alignment or using **infernal**.

⁹ The alignment of Fig. 3 can be downloaded from <http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/07-011/>.

Fig. 3. Alignment (top) and consensus secondary structure (below) of vertebrate vRNAs. The two internal polIII promoter sequences, Box A and Box B as well as the terminator sequence CTTT are highlighted. The ENSEMBL vRNA predictions end before the terminator motif, which strictly speaking it does not belong to the vRNA sequence, but is indicative of a true polIII transcript. In the secondary structure below the variable region has been left out (denote by three shaded gap symbols). The IDs of the experimentally known vRNA sequences are marked in green, novel candidate sequences from chicken and teleosts are highlighted in red. The remaining sequences can also be found with blast using one of the known sequences as query.

4.3 Telomerase RNAs

Telomerase maintains telomere length by adding G-rich telomeric repeats to the ends of linear eukaryotic chromosomes. The core telomerase enzyme consists of two components: an essential RNA component, which also serves as template for the repeat sequence, and the catalytic protein component TERT. The RNA component varies dramatically in sequence composition and in size. Although dozens of tel-RNAs (usually called TER in vertebrates and TLC-1 in yeasts) have been cloned and sequenced in the last years, the known examples are restricted to three narrow phylogenetic groups: vertebrates, yeasts, and ciliates. Although there appears to be a common core structure of all these tel-RNAs [21], and despite their length, tel-RNAs remain a worst case scenario for homology search.

Indeed, a survey of vertebrate tel-RNAs [4] shows dramatic sequence variation with only a few, short, well-conserved sequence patterns separated by regions of highly variable length. Some vertebrate groups, however, are still not represented. For example, blast-based approaches were unsuccessful in finding tel-RNAs in the genomes of teleost fishes. Using fragrep2, however, we were

able to find candidates in all five available teleost genomes and to verify their function experimentally. These results are described in a separate publication [6].

Yeast tel-RNAs appear to be even less well conserved: In [22], only seven short sequence motifs are reported in the more than 1.2kb transcripts of *Kluyveromyces* species, and of these only a few are partially conserved in *Saccharomyces*. In fact, *Saccharomyces* and *Kluyveromyces* TLC-1 genes cannot be aligned by programs such as `clustalw`. As a consequence search patterns for `fragrep2` have to be obtained manually. The search for tel-RNAs in a diverse set of fungal species is ongoing research, on which we will report elsewhere.

5 Concluding Remarks

We developed `fragrep2` to overcome limitations in the `fragrep` program [5] that became apparent in several applications: First, sequence patterns are rarely so well conserved that a classical IUPAC consensus sequence is a good descriptor. Hence we use PWMs. Secondly, in contrast to features such as transcription factor binding sites, even the conserved regions of ncRNAs seems to allow occasional indels. Hence we allow indels when matching PWMs to the genome. Thirdly, entire domains that are well conserved in wide range of clades are sometimes lost or mutated beyond recognition. Hence we allow the deletion of entire PWMs. As a result we obtain a tool, `fragrep2`, which is sensitive enough to find distant homologs that are not detectable with `blastn` while at the same time being fast enough to scan mammalian genomes even at the stage of shotgun reads. In the case of teleost fish telomerase RNA, we successfully characterized several predicted sequences experimentally [6].

Of course `fragrep2` also has its limitations. In particular, if the individual PWM patterns become small and ambiguous, and if the lengths of the intervals separating them become very variable, `fragrep2` becomes slow and, like all homology search tool operating at the limits of their sensitivity, tends to produce many false positives. Thus we are currently exploring the possibility to combine `fragrep2` with other homology search tools, either as a pre-processor for structure-based approaches, or as a post-processor for fast, regular-expression filters.

Availability: An implementation of `fragrep2` can be downloaded from <http://www.bioinf.uni-leipzig.de/Software/fragrep>.

References

1. Lowe, T., Eddy, S.: tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* **25** (1997) 955–964
2. Nawrocki, E.P., Eddy, S.R.: Query-dependent banding for faster RNA similarity searches. *PLoS Comp. Biol.* **3** (2007) e56 doi:10.1371/journal.pcbi.0030056.
3. Weinberg, Z., Ruzzo, W.L.R.: Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics* **22** (2006) 35–39

4. Chen, J.L., Blasco, M.A., Greider, C.W.: Secondary structure of vertebrate telomerase RNA. *Cell* **100** (2000) 503–514
5. Mosig, A., Sameith, K., Stadler, P.F.: **fragrep**: Efficient search for fragmented patterns in genomic sequences. *Geno. Prot. Bioinfo.* **4** (2005) 56–60
6. Xie, M., Mosig, A., Qi, X., Li, Y., Stadler, P.F., Chen, J.L.: Structure and function of the smallest vertebrate telomerase RNA from teleost fish. in preparation.
7. Kel, A.E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., E., W.: **MATCHTM**: a tool for searching transcription factor binding sites in DNA sequences. *Nucl. Acids Res.* **31** (2003) 3576–3579
8. Dinkelbach, W.: On nonlinear fractional programming. *Manage. Sci.* **13** (1967) 492–498
9. Schaible, S.: Fractional programming. *Z. Operations Res.* **27** (1983) 39–54
10. Arslan, A.N., Egecioglu, Ö.: Efficient algorithms for normalized edit distance. *J. Discr. Algorithms* **1** (2000) 3–20
11. Arslan, A.N., Egecioglu, Ö., Pevzner, P.: A new approach to sequence comparison: Normalized sequence alignment. *Bioinformatics* **17** (2001) 327–337
12. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the aminoacid sequences of two proteins. *J. Mol. Biol.* **48** (1970) 443–452
13. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A.: **Rfam**: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33** (2005) D121–D124
14. Piccinelli, P., Rosenblad, M.A., Samuelsson, T.: Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res.* **33** (2005) 4485–4495
15. Woodhams, M.D., Stadler, P.F., Penny, D., Collins, L.J.: RNase MRP and the RNA processing cascade in the eukaryotic ancestor. *BMC Evol. Biol.* **7** (2007) S13
16. van Zon, A., Mossink, M.H., Scheper, R.J., Sonneveld, P., Wiemer, E.A.C.: The vault complex. *Cell. Mol. Life Sci.* **60** (2003) 1828–1837
17. van Zon, A., Mossink, M.H., Schoester, M., Scheffer, G.L., Scheper, R.J., Sonneveld, P., Wiemer, E.A.C.: Multiple human vault RNAs. *J. Biol. Chem.* **276** (2001) 37715–37721
18. Kickhoefer, V.A., Searles, R.P., Kedersha, N.L., Garber, M.E., Johnson, D.L., Rome, L.H.: Vault ribonucleoprotein particles from rat and bullfrog contain a related small RNA that is transcribed by RNA polymerase III. *J Biol Chem* **268** (1993) 7868–7873
19. Vilalta, A., Kickhoefer, V.A., Rome, L.H., Johnson, D.L.: The rat vault RNA gene contains a unique RNA polymerase III promoter composed of both external and internal elements that function synergistically. *J Biol Chem.* **269** (1994) 29752–29759
20. Kickhoefer, V.A., Emre, N., Stephen, A.G., Poderycki, M.J., Rome, L.H.: Identification of conserved vault RNA expression elements and a non-expressed mouse vault RNA gene. *Gene* **309** (2003) 65–70
21. Chen, J.L., Greider, C.W.: An emerging consensus for telomerase rna structure. *Proc Natl Acad Sci U S A* **101**(41) (Oct 2004) 14683–14684
22. Tzfati, Y., Knight, Z., Roy, J., Blackburn, E.H.: A novel pseudoknot element is essential for the action of a yeast telomerase. *Genes & Dev.* **17** (2003) 1779–1788