Published online 27 September 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 18* ***6013–6020***
doi:10.1093/nar/gkn578

KnotInFrame: prediction of −1 ribosomal frameshift events

Corinna Theis, Jens Reeder and Robert Giegerich*

Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany

Received June 12, 2008; Revised and Accepted August 28, 2008

ABSTRACT

Programmed −1 ribosomal frameshift (−1 PRF) allows for alternative reading frames within one mRNA. First found in several viruses, it is now believed to exist in all kingdoms of life. Strong stimulators for −1 PRF are a heptameric slippery site and an RNA pseudoknot. Here, we present a new algorithm *KnotInFrame*, for the automatic detection of −1 PRF signals from genomic sequences. It finds the frameshifting stimulators by means of a specialized RNA-pseudoknot folding program, fast enough for genome-wide analyses. Evaluations on known −1 PRF signals demonstrate a high sensitivity.

INTRODUCTION

In the middle of the last century, George Beadle and Edward Tatum (1) proposed the 'one gene–one enzyme' theory, stating that one gene holds the genetic information of one enzyme, performing one enzymatic reaction. It was later changed to 'one gene–one polyprotein' to account for the discovery of polyproteins. However, today there are several cases which violate this assumption. By alternative splicing several protein variants are made by selecting different sets of exons from one single gene. With *trans*-splicing, proteins are synthesized from exons of different genes (2). Another way of enhancing the information content of an mRNA are translational recoding events. With codon hopping (bypassing) (3), the ribosome skips over a gap in the open reading frame (ORF) and resumes translation at a downstream codon. In procaryotes, UGA codons followed by a characteristic stem–loop structure are redefined from stop codons to insertion of selenocystein (4).

Programmed ribosomal frameshift (PRF), examined in this work, alters the reading frame by shifting the ribosome exactly 1 nt to either the +1 or −1 direction. A stop codon in the original reading frame is then bypassed in the shifted reading frame. This way, two different protein products can be obtained from one mRNA. The frequency of frameshifting events at a particular site is used by viruses for a defined ratio between the two proteins. Changes in the ratio can lead to less efficient virus propagation and thus, can be a target for antiviral therapeutics. Recently, a role of −1 PRF in posttranscriptional regulation has been proposed in *Saccharomyces cerevisiae* (5). The authors propose a model where −1 PRF leads to premature termination targeting the mRNA for rapid degradation via the nonsense-mediated mRNA decay pathway.

When two *cis*-acting signals are present, −1 PRF is most effective: a heptameric *slippery site* and a stable secondary structure. The slippery site is the location of the actual frameshift event having the consensus sequence X XXY YYZ (triplets are shown for the preshifted reading frame). The structural element follows within a few bases and can be a simple stem–loop structure or a pseudoknot. It is believed that the pseudoknot promotes a higher frameshift efficiency, since it is more effective in pausing the ribosome. It is also known that ribosome pausing is a necessary, but not sufficient prerequisite for frameshifting (6). The exact mechanism, by which the ribosome is brought into a different reading frame is still unknown. Two different models were recently proposed in (7) and (8).

PREVIOUS WORK

Several computational studies, with the general goal of detecting new PRF events, were undertaken in recent years (9–11,5). The first study by Hammell *et al.* (9) found over 200-putative PRF events in the yeast genome. Their approach relies on a strict pseudoknot structure consensus and requires two overlapping ORFs of at least 50 codons. In (11), a machine learning approach is used to discriminate between strong- and weak-PRF signals. Another approach (10) uses a combinatorial folding routine to identify possible frameshifting pseudoknots next to a slippery sequence.

The most recent and probably most elaborate study has been used to identify over a 1000 strong and statistically significant −1 PRF signals in the genome of

*To whom correspondence should be addressed. Tel: +49 521 106 2903; Fax: +49 521 106 6411; Email: robert@techfak.uni-bielefeld.de

S. cerevisiae (5). It is based on a two-step procedure, where the first step identifies slippery sequences followed by a possible frameshift pseudoknot. In the second step, candidates are analyzed statistically. In more detail, *RNAmotif* (12) is used in the first step and finds all potential pseudoknots that could be formed according to a descriptor. The descriptor specifies the allowed loop and helix lengths which are extracted from known −1 PRF pseudoknots. All potential pseudoknots are then subjects of statistical analysis in step two. The potential pseudoknots are refolded with *pknots* (13), a minimum free energy (MFE)-based RNA folding algorithm, capable of predicting a wide class of pseudoknots. The MFE of the folding is compared to folding energies of randomized sequences via *z*-score analysis. Finally, sequences with a low *z*-score (<-1.65) are regarded as statistically significant, since they appear to be more stable than expected by random.

This is the work we build upon. Despite its overall good architecture, there is one shortcoming in the procedure. Sequences are forced to fold into a pseudoknot in the first step but are folded freely in the second step. Thus, the result of the procedure is indeed 2-fold: (i) the final candidates have the theoretical potential to fold a pseudoknot, and (ii) they have a MFE folding which is more stable than expected by random. However, it is not guaranteed that the stable structure which causes the good *z*-score actually is a pseudoknot. In fact, we found that from 1679 strong candidates only 163 contain a pseudoknot. The pseudoknot, which was folded by *RNAmotif* in step one, may have an energy similar to the MFE folding, but more likely, it will be less stable and hence, less probable to be formed in equilibrium.

In the light of these findings, we propose to avoid a purely combinatorial matching step, such as *RNAmotif*. We develop a specialized RNA-folding program, called *pknotsRG-fs*, which explicitly folds a given sequence into the most stable structure conforming to the general frameshifting pseudoknot restrictions. The restricted folding energy can then be compared with the unrestricted MFE folding. This already gives a strong indication of whether the frameshift signal is likely to form or not.

In the following sections, we will first introduce the specialized folding program and show its incorporation into a tool for genome-wide annotation of −1 PRF signals. We conclude with an evaluation and a detailed comparison to the findings of (5).

MATERIALS AND METHODS

Building a −1 PRF consensus

The RECODE database (14) is the prime resource for information about translational recoding events. It contains almost a 100 −1 PRF recoding events in total, of which 28 entries are annotated with a 3′-RNA pseudoknot as *cis*-element. These sequences have been derived from different sources such as eukaryotic viruses, eukaryotes and bacteria.

We analyzed all available sequence and structure information in order to extract the consensus information displayed in Figure 1. The slippery sequence has the consensus X XXY YYZ, where XXX stands for any three identical nucleotides, YYY for either three As or three T/Us and Z for any nucleotide. The spacer region must contain at least 1 nt and not more than 12 nt. The stem and loop regions are chosen to include all but two examples which have unusually large loop sizes (see Supplementary Table S1). It should be noted that loops 2 and 3 are large enough to further fold internally and thus, can have a stabilizing effect on the overall structure. For example, the pseudoknot of the SARS coronavirus frameshift signal contains an additional third stem in loop 3 (15).

pknotsRG-fs: a specialized RNA-folding program

Following these observations, we need a specialized RNA-folding algorithm, computing the structure of lowest free energy while maintaining the given restrictions. Finding the structure with optimal energy can in general be done with dynamic programming (DP) algorithms, such as

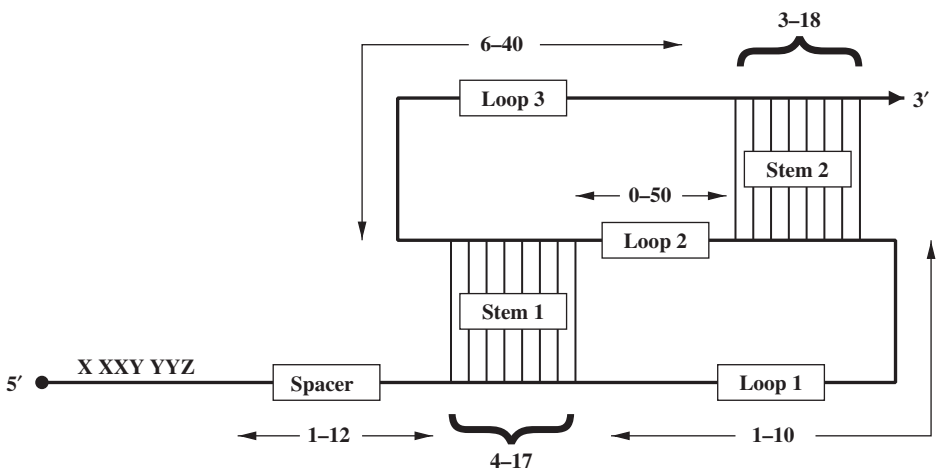

Figure 1. The consensus −PRF signal derived from the RECODE database.

Mfold (16) or *RNAfold* (17) for unknotted structures. However, explicitly formulating the exact recurrences which evaluate the folding space according to the above restrictions is a difficult and error-prone task. A remedy to this is to use a more high-level approach: the RNA-pseudoknot folding program *pknotsRG* (18,19) describes its search space by means of a grammar. The underlying thermodynamic model (20) (also used by *Mfold* and *RNAfold*), upon which optimizations takes place, is encapsulated in an 'evaluation algebra'. This concept of describing a DP by a grammar and an evaluation algebra has been formalized in the *algebraic dynamic programming* approach (ADP) (21–23).

By certain changes to the original *pknotsRG* grammar, we obtain a new grammar precisely describing frameshift-inducing pseudoknots: we incorporate all explicit length constraints displayed in Figure 1, but leave enough freedom for the loop regions to fold into any further stabilizing structures. Furthermore, the modified grammar requires stem 1 of the pseudoknot to start within 1–12 bases from the head of the sequence; this accommodates the spacer. The underlying thermodynamic model then comes for free—it has already been implemented with *pknotsRG* and can be re-used. We eventually compile the grammar into a low-level programming language (C) using the ADP compiler (24). This compilation creates and implements the explicit DP recurrences that are traditionally programmed by hand.

In this way, we obtain a specialized RNA-folding program *pknotsRG-fs*, which computes for a given input sequence the minimal free energy −1 PRF pseudoknot. Such a program has been called a thermodynamic matcher (TDM) in (18). If a sequence cannot form a pseudoknot with negative energy, the open structure is returned (with energy equal to zero). The compiled code takes less than a second to fold a sequence of length 100. The complete grammar is available at the project website. Readers with alternative ideas about the structure of the −1 PRF signal can use it as a starting point for a faithful implementation of a corresponding TDM.

We now explain how this program is incorporated in the −1 PRF prediction pipeline *KnotInFrame*.

KnotInFrame — a −1 PRF prediction pipeline

Overview. The −1 PRF prediction pipeline *KnotInFrame* is composed of three consecutive steps:

(i) In the *search phase*, we scan the input sequences for occurrences of the consensus slippery site in the correct reading frame. The downstream region of each slippery site is checked for suitability as a frameshift signal. This is done by comparing the minimal free energy of an enforced pseudo-knotted folding with the MFE of a freely folded structure. The former energy is computed by *pknotsRG-fs*, the latter by *RNAfold*.

(ii) In the *filtering phase*, three criteria based on the energy values of the free and the constrained folding are applied to reduce the number of candidates.

(iii) In the *ranking phase*, the candidates passing all filters are ranked by an evaluation function based on the normalized dominance of the pseudoknot.

Figure 2 gives an overview of the pipeline. Overall, *KnotInFrame* expects a set of sequences as input, and returns one or several predicted frameshift sites for each sequence. The predictions can then undergo comparative sequence analysis or experimental validation.

Definitions. Let s be a DNA sequence of length n. We require neither a start codon nor frame information. $s[i, j]$ is a substring of s starting at position i and ending at position j, and $s[i, n]$ is called a suffix of s.

A slippery sequence in s is a substring $s[i, i+6]$ that matches the consensus slippery motif of the form X XXY YYZ, where the spacer marks the zero-frame codons, XXX stands for any three identical nucleotides, YYY for either three As or three T/Us and Z for any nucleotide.

Given a slippery sequence p in s, a candidate x is a substring of s of the form $x = pu$, for various lengths of u.

Scanning for slippery sites. A forward scan of s yields a set P of putative slippery sites. P may be quite large, as the significance of the slippery motif is moderate—on a sequence with 50% GC content we expect about 8.8 occurrences per kilobase pair. P is reduced by *ISF*, the inframe stopcodon filter. The reduction of P is realized by scanning backwards from each $p \in P$, discarding p if an inframe stop codon is found before a start codon has been seen. Also, p is discarded if too close to the end of the sequence to accommodate the pseudoknot.

Folding of downstream regions with pknotsRG-fs and RNAfold. For each $p \in P$, we generate five candidate sequences $pu_k = s[i, i + |p| + 20 * k - 1]$ for $1, 2, 3, 4, 5$ and 6. This accounts for the potential pseudoknot sizes, which range in length from 26 to 118 in the RECODE database. For each candidate $x = pu$, we compute the MFE values for the best pseudoknot consistent with our consensus and the best unconstrained folding by calls of *pknotsRG-fs*(u) and *RNAfold*(u). Candidates now take the form $x = (pu, pknotsRG\text{-}fs(u), RNAfold(u))$.

Filtering of unlikely foldings. Subsequent filtering is based on the candidates' energy values. The low energy filter (LEF) discards candidates $x = pu$ where *pknotsRG-fs*(u) $> \alpha$. We choose the threshold $\alpha = -7.4$ kcal/mol, since all pseudoknots in our test set achieve this or a lower energy value. Next, the energy difference filter (EDF) discards candidates that rather fold into an unknotted structure, i.e. $RNAfold(u) + \beta < pknotsRG\text{-}fs(u)$. The default value for the parameter β is 8.7 kcal/mol, also derived from RECODE analysis.

The resulting set of candidates may still hold several predictions for the same slippery site. The normalized dominance filter (NDF) computes for each $x = pu$ the *length-normalized energy dominance*

$$\Delta(u) = \frac{RNAfold(u) - pknotsRG\text{-}fs(u)}{|u|} \quad \mathbf{1}$$

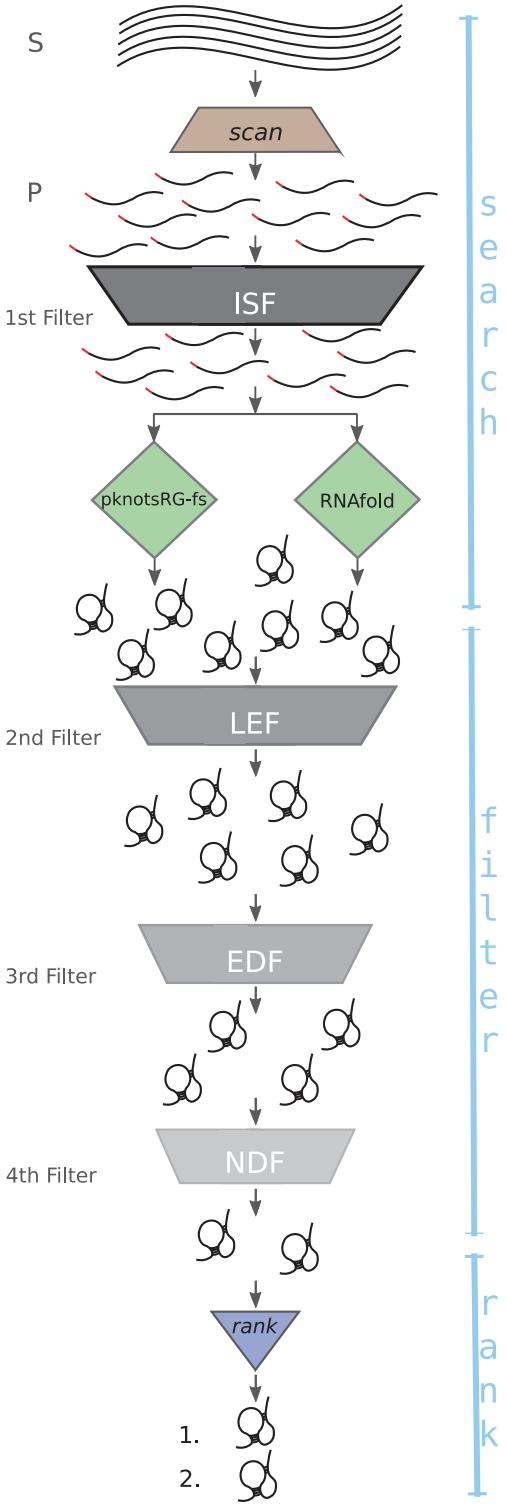

Figure 2. A diagram of the prediction pipeline *KnotInFrame*. In the search phase, a set of input sequences is searched for consensus slippery motifs. Discarding untranslatable sites by means of the first filter ISF and folding the remaining candidates with *pknotsRG-fs* and *RNAfold* flows into the *filter* phase. Three filters (LEF, EDF and NDF) discard further slippery sites based on the candidates' energy values. In a last phase *rank* the candidates will be ranked according to their normalized energy dominance.

Δ gives an indication of the stability of a structure, i.e. how strong a structure outweighs the other referred to their energy values. A positive Δ says that the pseudoknotted structure is more stable than the free-folded structure. For each p, the NDF retains only the candidate pu which maximizes $\Delta(u)$.

Score computation and ranking. In the last step of the pipeline, all remaining candidates are ranked in descending order of their $\Delta(u)$ values. Usually, we let the program report only the best (say 10) predictions per sequence. The final result of the pipeline is a list of the strongest frameshifting sites, their respective slippery site, the structural element and the free energy values leading to the ranking.

RESULTS

Evaluation on RECODE database

Here, we report on the evaluation of *KnotInFrame* on the RECODE DB. Since we built our consensus based on the structures stored in RECODE, this evaluation does not measure the correctness of the specialized folding program (the TDM). In fact, we take the correctness of the TDM for granted, based on our experience with implementing TDMs in ADP. What is really tested in the following is the adequacy of our scoring and ranking criterion—the normalized dominance. Naturally, our score is influenced by the thermodynamic parameters currently used. It might improve further if better thermodynamic measurements will be available for pseudoknotted structures.

We assume that the −1 PRF sites stored in the RECODE DB are true positives (TPs), which means a frameshift actually happens and the inducing structural element is a pseudoknot. Now, the questions we strive to answer are: (i) does *KnotInFrame* predict these TPs at all, (ii) on which rank does *KnotInFrame* predict the test candidates and (iii) what are the reasons for not finding or low ranking of TPs?

Table 1 displays the rank of our prediction for the annotated pseudoknots of the RECODE DB. Altogether, our pipeline predicts the real pseudoknots 24 times within the first five ranks: 17 times on rank 1, 4 times on rank 2 and once on rank 3. This demonstrates that *KnotInFrame* gives strong hints for frameshift sites.

Next, we have a closer look at the normalized dominance of the annotated pseudoknots and in particular the false negative (FN) cases, where *KnotInFrame* does not find the annotated positions or assigns to them a low rank.

Table 1. RECODE rank table

Rank	1	2	3	4	5	6	7	8	9	–
Frequency	17	4	1	0	2	0	1	0	1	2

This table sums up on which rank of our prediction we found the annotated pseudoknot of the RECODE DB that holds 28 −1 PRF entries with a pseudoknot as *cis*-element.
[a]The relevant frameshift site could not be predicted by *KnotInFrame*.

From Table 2, one can observe, that 20 of 26 real pseudoknots have a $\Delta > 0$. Sixteen of these 20 have also been predicted with our pipeline on rank 1. This, of course, confirms that a positive Δ is a strong indicator for frameshift sites. From our small RECODE test set it is hard to conclude a definite threshold for a good separation of strong and weak candidates. In Figure 3, we show that in fact true PRF signals tend to have a higher Δ than nonshifting slippery sites. However, the separation of the distributions is by no means sufficient for a clear classification. As a rule of thumb, we can state that candidates with $\Delta \geq 0.1$ are most likely true signals. With $\Delta \geq 0.05$, we capture approximately the same number of false positive (FPs) and TPs. Finding the appropriate balance between high sensitivity and selectivity for this problem should therefore be governed by the intended use of the program. We also note that there are annotated pseudoknots with a $\Delta \leq 0$. These are also the candidates where *KnotInFrame* failed to rank the true −1 PRF signal on a high position. We examined those FN in more detail and give explanations for their mispredictions:

- The dominance of is_m_is_is3 is undefined ('b' in Table 2) since the slippery site is out of *KnotInFrames* scope due to the remarkable appearance of the slippery sequence: it has a length of only 4 nt.
- The −1 PRF signal of pol_m_vir_gill has a total length of 177 nt but the longest substring the pipeline folds is 120 nt. A more detailed look at the annotation, reveals that the classification as a pseudoknot is at least questionable: the −1 PRF pseudoknot signal has been predicted by a hairpin folding via *Mfold*. Afterwards, the hairpin has been manually extended to a pseudoknot, by looking downstream for a complementary site which is able to pair with the hairpin loop region. This is also the reason for the exceptional length of the structure. An examination of the second stem with *RNAduplex* from the Vienna RNA package revealed a free energy of only −5.7 kcal/mol. In accordance with the energy models implemented in state-of-the-art pseudoknot folding programs, this energy might be too low to compensate for the destabilizing effects of the pseudoknot loops.
- A similar consideration holds for the entry of pol_m_vir_hcv. Its annotated pseudoknot is much longer than the maximal length of our consensus model. In this case, our pipeline folds at most the first 120 nt of the whole structure and hence can only find an alternative, suboptimal pseudoknot. Thus, this site is ranked low in our evaluation.
- A possible reason for predicting the frameshift signal of edr_m_euk_mmus on rank 5 might be that the structural frameshift component is not the pseudoknot reported in the RECODE. In (25) a more complicated pseudoknot with a 3 nt bulge in the first stem is reported. Our consensus model does not account for such pseudoknots and hence an alternative,

Table 2. Detailed analysis of normalized dominance for the RECODE test set

Organism (Abbreviation)	A annotated pseudoknot	A pseudoknot 1. Rank	annotated pseudoknot on Rank
pol_m_vir_hastr	0.070	–[a]	1
pol_m_vir_mhv	0.053	–[a]	1
edr_m_euk_mmus	−0.060	−0.015	5
pol_m_vir_eiav	0.040	0.133	5
pol_m_vir_fiv	0.068	–[a]	1
pol_m_vir_giar	−0.007	–[a]	1
pol_m_vir_hcv	−0.029	0.06	7
pol_m_vir_la	0.055	0.098	2
pol_m_vir_rsv	−0.023	−0.006	2
pol_m_euk_sars	0.085	–[a]	1
pol_m_vir_visna	0.083	–[a]	1
pol_m_vir_mmtv	0.180	–[a]	1
pol_m_vir_mpmv	0.050	–[a]	1
pol_m_vir_srv1	0.065	0.105	2
pol_m_vir_srv2	0.050	–[a]	1
edr_m_euk_hsap	0.057	–[a]	1
pol_m_vir_bev	0.122	–[a]	1
pol_m_vir_gill	–[b]	0.075	–[c]
pol_m_vir_porc	−0.025	0.024	9
pol_m_vir_cabyv	−0.002	0.065	3
is_m_is_is3	–[b]	−0.013	–[c]
pol_m_vir_ibv	0.045	–[a]	1
pol_m_vir_potato2	0.010	–[a]	1
pol_m_vir_potato1	0.038	–[a]	1
pol_m_vir_potato3	0.045	0.062	2
pol_m_vir_bwyv	0.150	–[a]	1
pol_m_vir_bydv1	0.035	–[a]	1
pol_m_vir_bydv2	0.163	–[a]	1

This table shows for 28 entries of RECODE DB, each with a pseudoknot inducing the −1 PRF, the normalized dominance (Δ) of the annotated pseudoknot and in comparison the A of the pseudoknot found on rank 1 with *KnotInFrame*. Additionally, the last column shows on which rank *KnotInFrame* predicts the real pseudoknot.
[a]Cases where the annotated and the predicted pseudoknot match. One can observe that 20 of 26 real pseudoknots have a $\Delta \geq 0.16$ of these 20 have been predicted on rank 1 with our pipeline. We conclude that a positive Δ gives a strong hint for a −1 PRF.
[b]There are no MFE results available, because the appropriate slippery site was not detected by *KnotInFrame*.
[c]The annotated structure was not found by *KnotInFrame*.

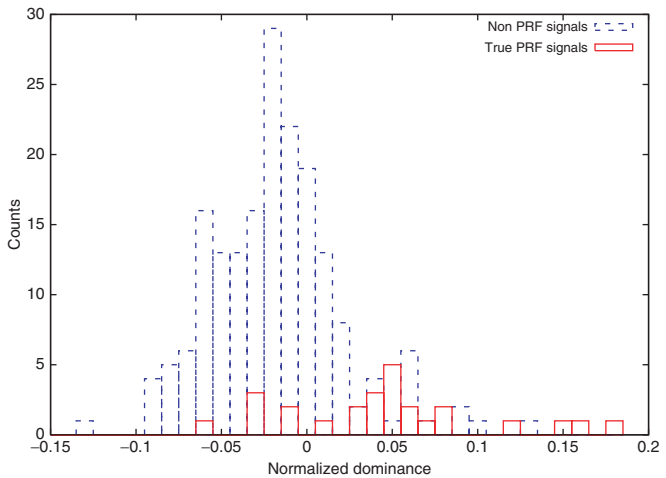

Figure 3. Histogram of the normalized dominance values (Δ) of 26 annotated pseudoknots in the RECODE database (red) against the normalized dominance of 183 slippery sites in our test set not leading to −1 PRF (blue).

suboptimal structure is folded. Consequently this structure is then ranked low by *KnotInFrame*.

Evaluation on frameshift signals from PseudoBase

We also tested *KnotInFrame* on frameshift examples not used for building the consensus, in order to test for any effects of overfitting. Unfortunately, there are only very few collected examples in public databases. However, we found seven −1 PRF signals not covered by RECODE in PseudoBase (26), a database designated to RNA pseudoknots of various biological functions. The corresponding genomic sequences sum up to 79 kb, which we analyzed with *KnotInFrame*.

The outcome is similar to the results obtained in the evaluation of RECODE. Five out of seven examples were predicted correctly on rank 1, another one was ranked second. Only one frameshift signal cannot be detected at all, since its slippery sites deviates from our consensus.

A note on *z*-score computation

At this point, the question arises if the true signals will also be ranked at a high position, when more slippery sites in a possibly longer input sequence compete for rank 1. A more detailed look (Figure 3) at the normalized dominance reveals that in fact, true −1 PRF signals have a tendency to have a higher Δ than the 183 slippery sites in our test set not leading to PRF (called FPs in the following). However, the discriminative signal is not strong enough for a clear separation.

In order to test the hypothesis that structural RNAs have a lower MFE than random sequences, we also performed a *z*-score analysis for all candidates reported by *KnotInFrame*. For each candidate, we shuffled 100 randomized sequences with the same dinucleotide content and folded them with *pknotsRG-fs*. In (27), it has been shown that using a TDM for *z*-score analysis provides a stronger signal than using a general folding algorithm. However, the outcome for −1 PRF signals is less fruitful than the results of (27) as shown in Figure 4. Most of the FPs have a *z*-score between 0 and −3. The fact that the *z*-scores are not centered at 0 is most certainly a result of *KnotInFrame* discarding unfavorable folds during the filtering phase. This already introduces a bias for more stable structures than expected by random. With a *z*-score of −3 and lower, the number of TPs equals approximately the number of true negatives (TNs). A small number of TPs is also deeply buried in the cloud of FPs which makes it hard to detect them either by *z*-score analysis or via the normalized dominance. Altogether, it turns out that with *z*-score analysis, we have the same difficulties in separating the true PRF signals from the wrong ones. Therefore, we refrain from using the *z*-score in *KnotInFrame*, since it would increase the runtime by a factor of 100 (for folding the randomized sequence) without a significant increase in performance.

It seems that with only one sequence at hand, we can do no better. However, the situation changes, if we have a set of homologous sequences. Then, the co-occurrence of a slippery site and a conserved, stable frameshifting pseudoknot could improve the search for −1 PRF signals in a future version of *KnotInFrame*.

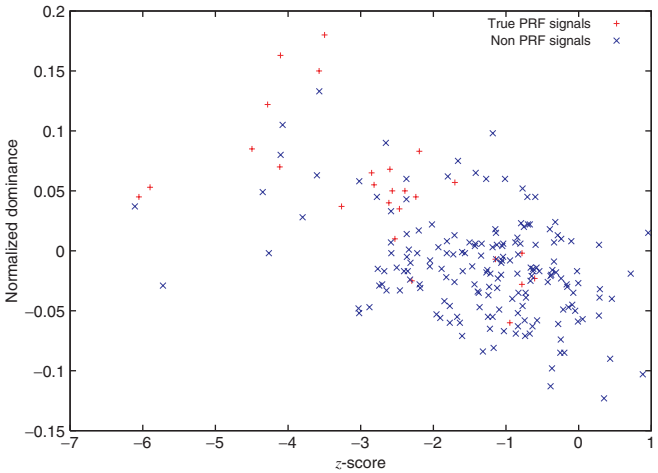

Figure 4. *z*-scores plotted against normalized dominance. Although only weakly correlated with Δ, the *z*-score analysis does not provide significant further information for separating −1 PRF signals from the background distribution.

Comparison with a previous yeast screen

We compared the predictions of *KnotInFrame* to the predictions of (5) on the *S. cerevisiae* genome. This comparison serves several purposes: since the reference set of verified signals in the RECODE DB is rather small, we use the predictions of (5) as extended test set. Of course, it is neither guaranteed nor expected that all of their predictions are in fact real frameshift events. However, a matching prediction of both approaches at a certain location strengthens the plausibility of this site. On the other hand, contradicting predictions may hint at a weak spot of either of the two approaches. We will show and discuss examples of each of these cases.

We downloaded the yeast genome sequences from NCBI (accessions: *NC_001133,...,NC_001148, NC_001224*; all versions are dated 24 January 2007) and extracted 6199 annotated ORFs with a total length of 8.76 MB. Each ORF was analyzed by *KnotInFrame* and for each ORF the five best predictions were stored and used in the comparison.

In the study of (5), 1679 candidates were classified as strong candidates, due to a good *z*-score and a low-MFE value. These candidates were made available by the authors in their PRFdb [(28), http://dinmanlab.umd.edu/prfdb/]. First, we asked what portion of those 1679 candidates were also predicted by our pipeline. On the first glance, the result was astonishing: only 257 sites were confirmed by our analysis, i.e. they were predicted within one of the first five ranks for each ORF. By visual examining of some of the missed candidates the reason was immediately obvious. The predicted structure in the PRFdb often does not resemble the pseudoknot frameshift consensus. Instead, we find virtually all possible other structures, such as a single hairpin, a chain of hairpins, bifurcated structures or complex pseudoknot structures. This supports our

main criticism of the approach of Jacobs *et al.*: the candidate sequences have the base-pairing potential to build a pseudoknot specified by the (*RNAmotif*) consensus, but for most sequences an alternative structure is more stable. This alternative structure is then used for the *z*-score computation and stored in the PRFdb. In contrast, our approach explicitly folds the consensus pseudoknot and the alternative structure and assigns to candidates with a more stable alternative structure a negative Δ and consequently a low rank. Therefore, we filtered the list of strong candidates in the PRFdb for structures containing at least one pseudoknot and obtained 163 candidates. Visual inspection revealed that this set still contains lots of structures neither conforming to our nor Jacobs *et al.*'s consensus. If we further filter out all structures violating the consensus (e.g. a too long spacer, too short helices, a too long loop 1, etc.; examples are provided in the Supplementary Data), we end up with only 74 structures — <5% of the original PRFdb. As expected, the overlap with our predictions is much larger for this set: 19 out of 74 were also predicted by *KnotInFrame*. In other words, the overlap for this set is around 26%, while for the complete PRFdb it is only 15%. We conclude that *KnotInFrame* supports the predictions of Jacobs *et al.* only partially, even if both methods were reduced to structurally similar putative PRF events.

Next, we tested if our predictions are supported by Jacobs *et al.* We created a list of our 100 strongest candidates by setting a threshold of $\Delta \geq 0.08$. Again, only a small portion (13) of these predictions are supported by the PRFdb. The reasons for this are manifold: first, Jacobs *et al.* are using a more stringent consensus, which allows only 1–3 bases in loop 1. Second, the folding program used in the creation of the PRFdb (*pknots*) uses an outdated energy model, while *KnotInFrame* uses the most up to date energy model for nested structures and an adapted model for pseudoknots. In consequence, it happens that a stable structure reported by *KnotInFrame* is rejected in the PRFdb due to a possibly worse energy reported by *pknots*. Remarkably, the maintainers of PRFdb started to include another RNA folding algorithm [*NUPACK* (29)] into the pipeline, which uses an energy model close to ours. Database entries derived from *NUPACK* analysis seem to be more consistent with our results. However, we cannot quantify the improvement, since the inclusion of *NUPACK* seems to be work in progress and only parts of the database are currently reprocessed.

Apparently, the method presented here and the one by Jacobs *et al.* differ substantially in their results. Various reasons have been given above, e.g. the differences in the structure prediction methods. However, we think the most influencing difference in both methods is the scoring system. While *KnotInFrame* employs normalized dominance, Jacobs *et al.* use *z*-scores and free energy for ranking. The fact that these two criteria are not equivalent can already be concluded from Figure 4, but has also been proven by our yeast screen comparison. In our opinion, normalized dominance is better suited for the task of discriminating true PRF signals from random ones. The *z*-score measures the evolutionary fitness of a sequence under evaluation against other (random) sequences that could have evolved instead, but were rejected during evolution due to an inferior fitness. In contrast, the normalized dominance makes an exact statement of how likely a given sequence folds into a frameshift inducing structure and not into an alternative one. In this sense, normalized dominance evaluates the 'fitness' of the PRF structure against its present competitors—those which exist in the folding space of the given RNA sequence, rather than against those that might exist in the past/future of a mutated sequence.

Run time analysis of *KnotInFrame*

To check the suitability of *KnotInFrame* for a large-scale application, we analyzed the run time of our tool which is basically effected by two factors: the folding routines and the number of slippery sites. The run time determining factor of the folding steps are the calls to *pknotsRG-fs*, hence we have to consider the time it needs to fold one sequence, which is on average 0.6 s for a substring of 120 nt. Note that we actually do not have to compute the foldings for the smaller subsequences $(40, \ldots, 100)$ separately. They can simply be backtraced from the dynamic programming matrix of the largest subsequence. Also, the theoretical run time depends on the probability of a slippery sequence to occur in a random sequence, given by:

$$P(\text{slippery site}) = 1 \times \underset{X}{\frac{1}{4}} \times \underset{X}{\frac{1}{4}} \times \underset{X}{\frac{1}{2}} \times \underset{Y}{\frac{1}{4}} \times \underset{Y}{\frac{1}{4}} \times \underset{Y}{1} \times \underset{Z}{1} = \frac{1}{512} = 0.195\%$$

Hence, the expected run time for an average sized bacterial genome of 5 MB with $5 \cdot 10^6 \times 0.00195 = 9750$ slippery sites (and thus 9750 calls of *pknotsRG-fs*) is $\sim 0.6\,\text{s} \times 9750 = 1\,\text{h}\ 37\,\text{min}$. In practice, the first filter (*ISF*) already discards a significant amount of slippery sites (42% for the RECODE test set), before the expensive folding step and run time decreases. This effect is somehow balanced by our observation that the real genomic sequences used in our evaluation contain more slippery sites than expected.

We measured the run time that *KnotInFrame* requires to analyse the 8.76 MB coding sequence of yeast. The whole computation with 43841 observed slippery sites (expected: 17109) can be performed in ~4.5 CPU hours, which is a definite improvement over previous approaches.

CONCLUSIONS

We presented *KnotInFrame*, a new and efficient tool for the automated detection of ribosomal −1 frameshift events. *KnotInFrame* employs a specialized RNA-folding program at its core to distinguish true −1 PRF events from random ones. The complete pipeline is fast enough to analyze complete genome sequences within a few hours. Our evaluation shows a high sensitivity on annotated −1 PRF events from the RECODE database. *KnotInFrame* clearly outperforms previous approaches in terms of compute resources. With our new method it is now possible to systematically annotate available genome

sequences. Positive results of such screens will help to further improve the tool's accuracy.

The *KnotInFrame* pipeline can be seen as a model case of developing a thermodynamic matcher, in our case *pknotsRG-fs*, and embedding it in a pipeline with further search criteria and filters. Development of thermodynamic matchers is supported by the tool *Locomotif* (30), which allows to graphically design RNA structures, annotated with sequence information and length restrictions. From such graphics, a thermodynamic matcher is produced automatically, bypassing tedious low-level programming and debugging. In the course of this work, we have also extended the repertoire of structural building blocks in *Locomotif* by a pseudoknot building block. In this way, it will be easy for workers in the field to construct their own pipelines akin to *KnotInFrame*, for RNA motifs of moderate size, with and without pseudoknots.

AVAILABILITY

KnotInFrame is available for online use at the Bielefeld Bioinformatics Server (BiBiServ) at http://bibiserv.techfak.uni-bielefeld.de/knotinframe. Also, we provide the underlying grammar of the folding program and detailed results of the yeast screen on the project's web site on the BiBiServ.

SUPPLEMENTARTY DATA

Supplemantary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Thomas Höchsmann for providing us with his *z*-score calculation scripts. We also want to thank the anonymous reviewer for pointing out to us several references which are not yet reflected by the annotations in RECODE.

FUNDING

Funding for open access charge: Bielefeld University.

Conflict of interest statement. None declared.

REFERENCES

1. Beadle,G.W. and Tatum,E.L. (1941) Genetic control of biochemical reactions in neurospora. *Proc. Natl Acad. Sci. USA*, **27**, 499–506.
2. Nilsen,T.W. (1993) Trans-splicing of nematode premessenger RNA. *Annu. Rev. Microbiol.*, 47, 413-440.
3. Herr,A.J., Atkins,J.F. and Gesteland,R.F. (2000) Coupling of open reading frames by translational bypassing. *Annu. Rev. Biochem.*, **69**, 343–372.
4. Walczak,R., Westhof,E., Carbon,P. and Krol,A. (1996) A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA*, **2**, 367–379.
5. Jacobs,J.L., Belew,A.T., Rakauskaite,R. and Dinman,J.D. (2007) Identification of functional, endogenous programmed −1 ribosomal frameshift signals in the genome of Saccharomyces cerevisiae. *Nucleic Acids Res.*, **35**, 165–174.
6. Kontos,H., Napthine,S. and Brierley,I. (2001) Ribosomal pausing at a frameshifter RNA pseudoknot is sensitive to reading phase but shows little correlation with frameshift efficiency. *Mol. Cell. Biol.*, **21**, 8657–8670.
7. Namy,O., Moran,S.J., Stuart,D.I., Gilbert,R.J.C. and Brierley,I. (2006) A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature*, **441**, 244–247.
8. Leger,M., Dulude,D., Steinberg,S.V. and Brakier-Gingras,L. (2007) The three transfer RNAs occupying the A, P and E sites on the ribosome are involved in viral programmed −1 ribosomal frameshift. *Nucleic Acids Res.*, **35**, 5581–5592.
9. Hammell,A.B., Taylor,R.C., Peltz,S.W. and Dinman,J.D. (1999) Identification of putative programmed −1 ribosomal frameshift signals in large DNA databases. *Genome Res.*, **9**, 417–427.
10. Moon,S., Byun,Y., Kim,H.-J., Jeong,S. and Han,K. (2004) Predicting genes expressed via −1 and +1 frameshifts. *Nucleic Acids Res.*, **32**, 4884–4892.
11. Bekaert,M., Bidou,L., Denise,A., Duchateau-Nguyen,G., Forest,J.-P., Froidevaux,C., Hatin,I., Rousset,J.-P. and Termier,M. (2003) Towards a computational model for −1 eukaryotic frameshifting sites. *Bioinformatics*, **19**, 327–335.
12. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
13. Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
14. Baranov,P.V., Gurvich,O.L., Hammer,A.W., Gestel,R.F. and Atkins,J.F. (2003) RECODE 2003. *Nucleic Acids Res.*, **31**, 87–89.
15. Plant,E.P., Prez-Alvarado,G.C., Jacobs,J.L., Mukhopadhyay,B., Hennig,M. and Dinman,J. D. (2005) A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.*, **3**, e172.
16. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
17. Hofacker,I., Fontana,W., Stadler,P., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
18. Reeder,J. and Giegerich,R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinfo.*, **5**, 104.
19. Reeder,J., Steffen,P. and Giegerich,R. (2007) pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids. Res.*, **35 (Suppl. 2)**, W320–W324.
20. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
21. Giegerich,R., Meyer,C. and Steffen,P. (2004) A discipline of dynamic programming over sequence data. *Sci. Comput. Program.*, **51**, 215–263.
22. Steffen,P. and Giegerich,R. (2005) Versatile and declarative dynamic programming using pair algebras. *BMC Bioinfo.*, **6**, 224.
23. Giegerich,R. and Steffen,P. (2006) Challenges in the compilation of a domain specific language for dynamic programming. In Haddad,H.M. (ed), *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*. Dijon, France, pp. 1603–1609.
24. Steffen,P. (2006) Compiling a domain specific language for dynamic programming. *PhD thesis*. University of Bielefeld, Germany.
25. Manktelow,E., Shigemoto,K. and Brierley,I. (2005) Characterization of the frameshift signal of Edr, a mammalian example of programmed −1 ribosomal frameshifting. *Nucleic Acids Res.*, **33**, 1553–1563.
26. vanBatenburg,F.H., Gultyaev,A.P. and Pleij,C.W. (2001) PseudoBase: structural information on RNA pseudoknots. *Nucleic Acids Res.*, **29**, 194–195.
27. Höchsmann,T., Höchsmann,M. and Giegerich,R. (2006) Thermodynamic matchers: strengthening the significance of RNA folding energies. In Markstein,P. and Xu,Y. (eds), *Comput. Syst. Bioinformatics. Conf.,* pp. 111–121.
28. Belew,A.T., Hepler,N.L., Jacobs,J.L. and Dinman,J.D. (2008) PRFdb: a database of computationally predicted eukaryotic programmed −1 ribosomal frameshift signals. *BMC Genomics*, **9**, 339.
29. Dirks,R. and Pierce,N.A. (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, **24**, 1664–1677.
30. Reeder,J., Reeder,J. and Giegerich,R. (2007) Locomotif: from graphical motif description to RNA motif search. *Bioinformatics*, **23**, i392–i400.