Searching the *Saccharomyces cerevisiae* genome for -1 frameshifting sites

Michaël Bekaert [1], Jean-Paul Forest[3], Laure Bidou[1], Alain Denise[3], Guillemette Duchateau-Nguyen[2,4], Céline Fabret[1], Christine Froidevaux[3], Isabelle Hatin[1], Jean-Pierre Rousset[1] and Michel Termier[2]

[1] Génétique Moléculaire de la Traduction, [2] Bioinformatique des Génomes, Institut de Génétique et Microbiologie (IGM), UMR CNRS 8621, [3] Laboratoire de Recherche en Informatique (LRI), UMR CNRS 8623, Université Paris-Sud, 91405 Orsay, France and [4] Hoffmann-La Roche Ltd, Basel, Switzerland

Introduction

In 1985, Jacks and Varmus described the first -1 ribosomal frameshifting, from which they established the canonical model of eukaryotic -1 frameshifting site (Jacks *et al*., 1985 and 1988). Today, several tens of viruses and one mouse nuclear gene (Shigemoto *et al*., 2001) have been identified as bearing such a -1 frameshifting site. A typical site contains a "slippery" heptamer in 5', where both A and P site tRNAs slip by one nucleotide upstream, followed by a stimulatory structure (stem loop, or pseudoknot) downstream (Brierley et *al*., 1989). The "slippery" heptamer is separated from the stimulatory structure by a short sequence (3 to 11 nucleotides), the so-called "spacer". Based on this model, several studies have been undertaken to identify frameshifting sites in the nuclear genome of the yeast *Saccharomyces cerevisiae* (Hammell *et al*., 1999 and Liphardt, 1999). However, none of these allowed to identify with certainty authentic expressed genes controlled by -1 frameshifting. Two reasons might be proposed: first, the model might not be precise enough, leading to the identification of too many false positive candidates (Bekaert *et al.*, 2003); conversely the model might be too rigid, failing to identify true positive candidates. This would be the case, for example, if -1 frameshift could be directed by a more "degenerated" structure, or by mechanisms that rely on other types of signals.

We used two independent strategies to look for frameshifting sites *in silico*.

- Similarity-based approach: we searched for genomic regions where two domains, each carrying a protein pattern, can be associated on a same polypeptide by a single -1 frameshifting. This approach does not rely on any model of frameshifting site.
- Model-based approach: we searched for genomic regions where a pseudoknot immediately follows a slippery sequence.

The two methods were then crossed and common hits evaluated *in vivo*.

Similarity-based approach

The first step of this approach was to seek genomic configurations compatible with a -1 ribosomal frameshifting by using the following criterion: two open reading frames, one in 0 frame (ORF0), the other in -1 frame (ORF-1), that overlap along an intermediate shared region. We fixed a length of at least 100 nucleotides for both ORF0 and ORF-1 areas, and of at least 150 base pairs for the whole structure.
All searches were performed independently on three sets of data: the *S. cerevisiae* genome (12Mbp), the genome of the yeast L-A virus (4579bp) (Icho *et al*., 1989) which is known to bear an authentic -1 ribosomal frameshifting site, and an artificial genome which exhibits the same hexamer frequencies as the yeast genome, using the GenRGenS software (Denise *et al*. 2003). From this analysis, 25,148 regions were found in the yeast genome, 21,135 in the artificial genome and 10 in the yeast L-A virus genome.

The second step consisted in filtering candidates by retaining regions that have protein patterns in both ORF0 and ORF-1. For this purpose, we used InterPro (release 5.3) (Apweiler *et al*., 2001), a database of protein families, domains and functional sites. Our approach was validated as far as only the actual frameshifting region was retrieved from the L-A virus genome. Moreover, we found 168 candidates in the yeast genome and 14 in the artificial genome. Among the 168 regions, three categories can be defined.

? ? The first one is characterised by domains that contain stretches of repeated amino-acids in each of the two frames (143 occurrences). Among them, 125 are DNA microsatellites (Hamada *et al*., 1984): tandem repetitions of the same triplet, which are read as repetitions of two different amino acids, depending on the reading frame. Noteworthy, such candidates were not found in the random genome.

? ? The second category is composed of regions where the two ORFs bear the same protein pattern, or two distinct but functionally compatible motifs (*e.g*. a sugar transporter and a sugar binding site). We found 12 such regions.

? ? The third category consists in regions which bear both functional regions and repetitions of amino acids. All the candidates from the random genome belong to this category.

Model-based approach

In the second approach, we designed an algorithm that finds potential frameshift sites according to the current model. First, the algorithm searches for a slippery sequence by means of a finite state automaton. Then it looks for a pseudoknot in the nucleotides that immediately follow the slippery sequence. Since we were looking only one simple kind of pseudoknots, general algorithms were not needed (*e.g*. Zuker *et al*., 1981, Rivas *et al.*, 1999). Thus, this second step simply consists in seeking consecutively two overlapping stems which could constitute a pseudoknot. To find each of these stems, the sequence is folded on itself so that the weighted sum of the base pairings is maximized (in a way similar to Nussinov *et al*., 1980). The score of each ordered pair (GC is different from CG) of nucleotides was assigned according to its frequency in 17 wild-type viral frameshift sites. The highest-scoring stem is kept only if its score is above a fixed threshold (which also depends on wild-type sites). Both the set of scores and the thresholds are distinct for the two stems. If a first stem is found, the 5'-arm of the second stem has to lie between the two arms of the first one. Constraints on the lengths of different stems and loops were added to ensure that the most unlikely candidates are discarded.
We checked that the algorithm folds all the known sites correctly, including the one in the L-A virus. It yielded 185 hits in the *S. cerevisiae* genome and 104 hits in the random genome. Both results were obtained in a couple of minutes.

Crossing the methods

Finally, we crossed both methods to find common candidates: by comparing the 168 regions of the first approach and the 185 of the second one, 4 common candidates were identified. As the two methods are completely independent, common candidates seem especially relevant.
The following step is to characterise these candidates. This will include verification of the sequence to eliminate sequencing errors, estimation of frameshifting efficiency and, possibly, biological characterisation.
All candidates were sequenced and were indeed in an overlapping configuration. For one of them, the overlap region was introduced into a dual-reporter vector (Bidou *et al*., 2000) in order to estimate *in vivo* the frameshifting frequency elicited by the candidate sequence. This candidate exhibits a frameshifting efficiency of 13%. Compared to the background rate (0.1%) and to the programmed yeast L-A virus frameshift (10%), this level is highly significant. Biological characterisation of this candidate is currently in progress. Other candidate regions are also being analysed.

Conclusion

The combination of two simple approaches thus made it possible to identify several candidate genes potentially controlled by a –1 frameshift mechanism. Our approach is promising and could be straightforwardly extended to other organisms, eukaryotic as well as prokaryotic (Bertrand *et al*., 2002).

Acknowledgements
This work was supported in part by a CNES grant on the PREMIER program and by a CNRS-INRA-INRIA-INSERM Bioinformatics grant.

References

Apweiler R., Attwood T.K., Bairoch A., Bateman A., Birney E., Biswas M., Bucher P., Cerutti L., Corpet F., Croning M.D.R., Durbin R., Falquet L., Fleischmann W., Gouzy J., Hermjakob H., Hulo N., Jonassen I., Kahn D., Kanapin A., Karavidopoulou Y., Lopez R., Marx B., Mulder N.J., Oinn T.M., Pagni M., Servant F., Sigrist C.J.A., Zdobnov E.M. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res., 29, 37-40.

Bekaert M., Bidou L., Denise A., Duchateau-Nguyen G., Forest J.-P., Froidevaux C., Hatin I., Rousset J.-P., Termier M. (2003). Towards a computational model for -1 eukaryotic frameshifting sites, Bioinformatics, 4, 2610-2621.

Bertrand C., Prère M.F., Gesteland R.F., Atkins J.A., Fayet O. (2002), Influence of the stacking potential of the base 3' of tandem shift codons on -1 ribosomal frameshifting used for gene expression, RNA, 8: 16-28.

Bidou L., Stahl G., Hatin I., Namy O., Rousset J.-P., Farabaugh P (2000) Nonsense-mediated decay mutants do not affect programmed -1 frameshifting. RNA, 6, 952-961.

Brierley I., Digard P, Inglis S.C. (1989) Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. Cell. 57, 537-547.

Denise A., Ponty Y., Termier M. (2003) Random generation of structured genomic sequences. Poster presented at RECOMB'03, Berlin, April 2003. http://www.lri.fr/~denise/GenRGenS.

Hamada H., Petrino M.G., Kakunaga T., Seidman M., Stollar B.D. (1984). Characterization of genomic poly(dT-dG).poly(dC-dA) sequences: structure, organization, and conformation, Mol Cell Biol, 19, 327-335.

Hammell A.B., Taylor R.C., Peltz S.W., Dinman J.D. (1999) Identification of putative programmed -1 ribosomal frameshift signals in large DNA databases. Genome Res., 9, 417-427.

Icho T., Wickner RB. (1989) The double-stranded RNA genome of yeast virus L-A encodes its own putative RNA polymerase by fusing two open reading frames. J. Biol. Chem., 264, 6716-6723

Jacks T., Varmus H. (1985) Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. Science, 230, 1237-1242.

Jacks T., Power M., Masiarz F., Luciw P., Barr P., Varmus H. (1988) Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature, 331, 280-283.

Liphardt J. (1999) The mechanism of -1 ribosomal frameshifting: experimental and theoretical analysis, PhD thesis, Churchill College, Cambridge.

Nussinov R., Jacobson A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci USA., 77, 6903-6913.

Rivas E ., Eddy S.R., (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol., 285, 2053-2068.

Shigemoto K., Brennan J., Walls E., Watson C.J., Stott D., Rigby P.W., Reith A.D. (2001) Identification and characterisation of a developmentally regulated mammalian gene that utilises -1 programmed ribosomal frameshifting. Nucleic Acids Res., 29, 4079-4088.

Zuker M., Stiegler P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res., 9, 133-148.