BIOINFORMATICS

Vol. 20 no. 2 2004, pages 155–160
DOI: 10.1093/bioinformatics/bth019

Transposable element annotation of the rice genome

Nikoleta Juretic[1,*], Thomas E. Bureau[1] and Richard M. Bruskiewich[2]

[1]Department of Biology, McGill University, Montreal, Quebec, H3A 1B1 Canada and
[2]Biometrics and Bioinformatics Unit, International Rice Research Institute,
DAPO Box 7777, Metro Manila, Philippines

Received on November 9, 2003; accepted on November 11, 2003

ABSTRACT

Motivation: The high content of repetitive sequences in the genomes of many higher eukaryotes renders the task of annotating them computationally intensive. Presently, the only widely accepted method of searching and annotating transposable elements (TEs) in large genomic sequences is the use of the RepeatMasker program, which identifies new copies of TEs by pairwise sequence comparisons with a library of known TEs. Profile hidden Markov models (HMMs) have been used successfully in discovering distant homologs of known proteins in large protein databases, but this approach has only rarely been applied to known model TE families in genomic DNA.

Results: We used a combination of computational approaches to annotate the TEs in the finished genome of *Oryza sativa* ssp. *japonica*. In this paper, we discuss the strengths and the weaknesses of the annotation methods used. These approaches included: the default configuration of RepeatMasker using cross_match, an implementation of the Smith–Waterman–Gotoh algorithm; RepeatMasker using WU-BLAST for similarity searching; and the HMMER package, used to search for TEs with profile HMMs. All the results were converted into GFF format and post-processed using a set of Perl scripts.

RepeatMasker was used in the case of most TE families. The WU-BLAST implementation of RepeatMasker was found to be manifold faster than cross_match with only a slight loss in sensitivity and was thus used to obtain the final set of data. HMMER was used in the annotation of the *Mutator*-like element (MULE) superfamily and the miniature inverted-repeat transposable element (MITE) polyphyletic group of families, for which large libraries of elements were available and which could be divided into well-defined families. The HMMER search algorithm was extremely slow for models over 1000 bp in length, so MULE families with members over 1000 bp long were processed with RepeatMasker instead. The main disadvantage of HMMER in this application is that, since it was developed with protein sequences in mind, it does not search the negative DNA strand. With the exception of TE families with essentially palindromic sequences, reverse complement models had to be created and run to compensate for this shortcoming. We conclude that a modification of RepeatMasker to incorporate libraries of profile HMMs in searches could improve the ability to detect degenerated copies of TEs.

Availability: The Perl scripts and TE sequences used in construction of the RepeatMasker library and the profile HMMs are available upon request.

Contact: njuret@po-box.mcgill.ca

INTRODUCTION

Domestic rice (*Oryza sativa*) is the world's most important crop and a model system for most cereal plants. In 1997, the International Rice Genome Sequencing Project (IRGSP) consisting of research groups from Japan, Peoples Republic of China, Thailand, Taiwan, Korea, France, Brazil, India, United Kingdom and United States was established with the goal of producing a high-quality rice genome sequence. The consortium used a clone-by-clone approach and a range of genetic resources, including a high-density physical map (Harushima *et al*., 1998), a large collection of expressed sequence tags (Yamamoto and Sasaki, 1997) and a yeast artificial chromosome-based physical map (Saji *et al*., 2001), was utilized in assembling the sequence. In December 2002, the IRGSP announced the completion of a high-quality draft sequence of the entire rice genome, with 10-fold coverage and 99.99% accuracy (http://www.tigr.org/new/press_release_12-18-02.shtml). The work described here was conducted as a part of the effort to analyze this sequence in detail.

Transposable elements (TEs) or transposons make up a large proportion of many eukaryotic genomes, and previous estimates of the TE content in the rice genome range from 10 to 25% (Mao *et al*., 2000; Turcotte *et al*., 2001). In addition, the TE complement of the rice genome is very diverse. The TE annotation of a 400 Mb genome with such a high proportion and diversity of TEs is a computationally challenging task that requires the application of different methodologies. We decided to evaluate some of the different methods available. In the genomics era, as the sequences of an increasing number of genomes are becoming available, it is important to

*To whom correspondence should be addressed.

establish the optimal approach to annotate these ubiquitous genomic residents.

The best known computational tool developed for the systematic genome annotation of repeat families is RepeatMasker (A.F.A.Smit and P.Green, unpublished data, http://ftp.genome.washington.edu/RM/RepeatMasker.html), which uses representative sequence libraries of known repeats to perform Smith–Waterman homology searches. However, eukaryotic genomes contain large amounts of transposon relics—ancient, highly degenerated TEs. The use of pairwise sequence comparisons will likely fail to detect these 'distant homologs' of known TE families.

McCarthy *et al*. (2002) developed LTR_STRUC, a program that uses structural features rather than sequence homology, to identify LTR retrotransposons in the rice genome. However, while this program is useful for finding previously unknown families, it is limited to detecting intact elements with well-conserved structural features and, thus, is suboptimal for genome annotation. RECON is a highly computationally intensive *de novo* approach capable of detecting and grouping repetitive sequences (Bao and Eddy, 2002). This approach could be very valuable for investigating the TE content of a genome that has not been studied previously, but with the large amount of information that already exists on rice repeat families, such a computational effort could be considered redundant.

Profile hidden Markov models (HMMs) are statistical models of multiple sequence alignments that have been used successfully to look for distant homologs of known protein families. Lazareva-Ulitsky and Haussler (1999) used an HMM-like method called dnA Sequence Alignment (ASA) for the alignment of the Alu-Sb subfamily of human SINE non-LTR retrotransposons, but to our knowledge this is the first application of profile HMMs in the search for members of known TE families on a genomic scale.

METHODS

Input data

IRGSP pseudomolecules Twelve reference molecules or pseudomolecules, corresponding to the 12 rice chromosomes, were assembled from 3157 clones (T.Sasaki, personal communication). The finished sequences of rice chromosomes 1 (Sasaki *et al*., 2002), 4 (Feng *et al*., 2002) and 10 (Rice Chromosome 10 Sequencing Consortium, 2003) have been previously reported. 155 physical gaps and 5963 sequence gaps remain in the 12 chromosomes. One hundred Ns were inserted to mark the physical gaps and 50 Ns for the sequence gaps. The total nucleotide sequence length of the pseudomolecules excluding the Ns is 357.6 Mb.

Rice transposable element database (RTEdb) The TIGR Rice Repeats Database (TIGR RDB, http://www.tigr.org/tdb/e2k1/osa1/) was used as the starting point. Published rice TE sequences that were absent from TIGR RDB were incorporated into RTEdb. Finally, several unpublished retrotransposons (C.Vitte, personal communication) and a large set of unpublished *Mutator*-like elements (MULEs; Bureau, personal communication) were also added. The completed database, in the form of a 10.6 MB text file, contained 18 342 complete and partial FASTA sequence elements.

Algorithms

RepeatMasker RepeatMasker is a widely known program that screens DNA sequences for interspersed repeats and low complexity sequences. The default setting of RepeatMasker is to run the program cross_match as the search engine. Cross_match is an implementation of the Smith–Waterman–Gotoh algorithm developed by Phil Green (unpublished data), and is highly sensitive but considerably slow when analyzing long sequences. MaskerAid is a script (Bedell *et al*., 2000, http://sapiens.wustl.edu/MaskerAid) that makes it possible to use WU-BLAST (http://blast.wustl.edu/, W. Gish, personal communication) instead of cross_match, thus increasing the speed of search. Although the sensitivity of MaskerAid is slightly inferior to that of cross_match, even when run in sensitive mode, the vast reduction in run time makes this option very attractive, especially when processing very long sequences such as pseudomolecules. Changing the speed settings when using MaskerAid has very little effect on the speed, while significantly changing sensitivity, thus we decided to use WU-BLAST in the sensitive mode for all the RepeatMasker runs. As a benchmark, RepeatMasker was run on the rice PAC clone P0035F12 (GenBank accession no. AP003313.3) using either cross_match or WU-BLAST and the output was compared. WU-BLAST detected 69/71 TEs found by cross_match, but it also showed one additional hit missed by cross_match. Since the execution time with WU-BLAST was 16 times shorter than with cross_match, it was selected as the optimal method under the circumstances. Checking for bacterial insertion sequences was needlessly time consuming so this step was skipped, as well as the step in which low complexity repeats are masked.

The ProcessRepeats script, which runs as a part of the RepeatMasker program, makes assumptions as to whether nearby fragments are a part of the same element, in which case it assigns them the same ID. This is an important feature of RepeatMasker, since different parts of the same element can often be separated by another nested element or degenerate internal sequence.

Profile HMMs Profile HMMs were originally developed as speech recognition tools, but in the last decade they have been adapted to detecting conserved patterns in multiple biological sequences (Haussler *et al*., 1993; Krogh *et al*., 1994). Profile HMMs capture position-specific information concerning the level of residue conservation, and the likelihood of

each residue to occur at that position. The typical profile HMM is a chain of match (M), insert (I) and delete (D) nodes, with all transitions between nodes and all character costs in the insert and match nodes trained to specific probabilities. It has been shown by Park *et al*. (1998) that profile-based methods are much more effective in detecting remote protein homologs than traditional pairwise methods, such as BLAST (Altschul, 1991) and FASTA (Pearson and Lipman, 1988). HMMs have had many applications, ranging from gene finding and detecting regulatory modules in genomic sequences to finding protein homologs in protein databases (Pedersen and Hein, 2003; Sinha *et al*., 2003). Although they were used much more extensively in a protein context, they have been frequently applied to DNA sequence as well, and some of the successful non-comparative gene finders, such as GENSCAN and HMMgene, are HMM-based.

HMMER (Eddy, 1998, http://hmmer.wustl.edu/) is an open-source package for building and searching with profile HMMs. Version 2.3 was used in this analysis to construct profile HMMs for families of TEs. HMMER uses a profile architecture called Plan 7, which is somewhat more complex than the basic profile HMM architecture. The core section of Plan 7, essentially the original Krogh/Haussler architecture, is composed of M/I/D nodes flanked by begin (B) and end (E) states, but it also has 'special' S, N, C, T and J states, and no D to I or I to D transitions. When using Plan 7, there are no traditional 'global' and 'local' alignments. Instead, all alignments are global, but most of the sequence may be assigned to N, C and J states that generate 'random' sequence. Multiple hits can be found in the same sequence, as defined by the E to J transition and the J state (Eddy, 2003).

Sequence Alignment and Modeling (SAM, Hughey and Krogh, 1996) software system is another popular open source package of profile HMM software for biological sequence analysis. The main difference between the two packages is how they represent local alignments as global. While HMMER uses the N, C and J states to represent unaligned flanking sequence, SAM uses free insertion modules (FIMs), which consist of D and I states only. They are used at the beginning and end of the model to allow an arbitrary number of insertions at either end.

Madera and Gough (2002) compared the performance of HMMER and SAM on two protein databases. They found that SAM produced better models, but HMMER was between one and three times faster when searching large databases. In addition, HMMER's model building was found to be less susceptible to error stemming from presence of 'poisoning sequences', i.e. sequences that contained parts of domains from other protein superfamilies. This is important when applied to TEs, since many TEs contain acquired host sequences or nested elements that can be considered as 'poisoning sequences'. For these reasons, we chose to use HMMER rather than SAM. Since the HMMER package does not include a multiple sequence alignment program, all alignments were performed with ClustalW (Thompson *et al*., 1994), and manually adjusted if deemed necessary.

When there is a high degree of redundancy in the set of sequences being used to construct a profile HMM, the model will over-represent the similar sequences, unless a weighting algorithm is applied to downweight counts from closely related sequences and upweight distantly related sequences. The default Gerstein/Sonnhammer/Chothia tree-weighting algorithm was used in this analysis, although several other weighting methods are available as options in HMMER (Eddy, 2003).

BLAST As a complement to HMMER and RepeatMasker searches, the sequences of predicted rice open reading frames (ORFs) were searched using BLASTP with a set of TE-related protein sequences.

Post-processing The results generated were parsed into ('Gene Finding Format' or 'General Feature Format' GFF), using Perl object modules developed at the Sanger Centre (http://www.sanger.ac.uk/Software/formats/GFF). While RepeatMasker output can be parsed directly into GFF with the –gff option, a Perl script was developed using a modification of HMMER parsing code from Bioperl (http://bioperl.org), to convert the HMMER raw output into GFF. Use of the simple tab delimited annotation record format of GFF facilitated manipulations, such as filtering, transforming and intersecting of data generated by different programs and statistical analysis using Windows-based programs such as Microsoft Excel.

All results obtained by either RepeatMasker or HMMER were merged using the intersect_overlap_merge method of the Sanger GFF::GeneFeatureSet.pm module. This step ensured that no region of genomic sequence was annotated as more than one element, which could have occurred in the case of elements nested within other elements or if, e.g. a region was annotated as belonging to two different MULE families due to their similar coding sequences. The merged results were used to calculate genome coverage, but the original counts were retained to avoid losing the nested elements.

Execution

All computations were performed on a Compaq AlphaServer DS20E dual processor EV667 with 1 GB of RAM running Tru64 UNIX, or on a Dell Precision 620 dual processor 1 GHz Intel Xeon processor with 1 GB of RAM running Red Hat 7. The versions of WU-BLAST used by MaskerAid were BLASTN2.0MP-WashU (18 January 2003; decunix5.0a) and (23 May 2003; linux24-i686), respectively.

Results of the TE annotation of the rice genome will be published as a part of the IRGSP rice genome paper.

RESULTS AND DISCUSSION

Once the RTEdb was complete, we attempted running RepeatMasker on the sequence of rice chromosome 10, using RTEdb

as the library. After more than 150 h, the run was still not completed, so we decided to explore alternative approaches.

In RTEdb, 15 064 of the 18 342 sequences were designated as various families of Miniature Inverted-repeat TEs (MITEs). MITEs are a polyphyletic group of elements that have family-specific terminal inverted repeats (TIRs), are small in size and occur at high copy number. MITEs most likely arose by internal deletions of autonomous class II elements, and various lines of evidence indicate that *Tourist* elements are related to the IS5/*Tourist* superfamily, while *Emigrant* and *Stowaway* are part of the IS*630*/Tc1/*mariner* superfamily of TEs (Turcotte *et al*., 2001), and should thus be classified as *mariner*-like elements. Apart from these three major groups, a variety of other TEs with uncertain relationships to known superfamilies have been labeled as MITEs. For simplicity, the name MITEs will be retained in the analysis to describe collectively all these groups. It was clear that the first step in accomplishing the annotation was the reduction in the effective number of sequences to be used as queries in the search. Profile HMMs represented a promising option since constructing a limited number of profiles from numerous individual sequences could result in a sharp decrease in search time complexity.

In order to determine how this approach would be compared with RepeatMasker in terms of sensitivity and time cost, a simple benchmarking experiment was conducted. BAC OSJNBa0078001 (GenBank accession no. AC079888.9), which was determined by a BLAST search to contain three *Stowaway* XIV elements, was used to assess the sensitivity. It was searched with RepeatMasker using the complete set of *Stowaway* XIV elements as the library, and with the HMMER program hmmsearch using a profile HMM built from an alignment of the same set of elements. RepeatMasker produced three hits above the default score threshold of 200, two of which corresponded to elements identified by BLAST. HMMER produced 10 hits overall, including all three identified by BLAST and seven partial elements, which had much lower scores. In conclusion, the HMMER search was at least as sensitive as RepeatMasker. To test how HMMER performs time-wise, the MITE adh-2 family, consisting of 78 members, was run on the chromosome 10 pseudomolecule. Both RepeatMasker and HMMER took 18 min to complete this run. However, when the *Stowaway* I family, with 686 intact members, was used instead, the RepeatMasker run lasted 40 min, while HMMER still finished in 18 min. The speed of execution of a computer program is relative, as it depends on the computer architecture used. Ideally, the quality of results should be the only factor considered when choosing a method; however, in practical situations the time component is rarely irrelevant.

Ninety-six different designations of MITE families were present in the database. Members of each of these families were aligned using ClustalW and a representative member was selected for an alignment of all the families. This global alignment showed that a number of groups with different names aligned very well, i.e. belonged to the same family. Merging these groups reduced the number of families from 96 to 64, including 26 *Tourist* and 20 *Stowaway* families.

Constructing the alignments was the most labor-intensive, but also the most important step of the analysis, since the quality of multiple sequence alignments used to build models is the most significant factor affecting the overall performance of profile HMMs (Madera and Gough, 2002). Forty-eight MITE families that contained at least three intact members were removed from the RTEdb. Intact members from each family were aligned and the resulting alignments were used to build local, multidomain profile HMMs. Since the elements in the database were not compiled in a particular orientation, each family had to be aligned a minimum of two times—the first alignment revealed which elements were present in the same orientation, so for the second alignment one of the two groups the reverse-complement was examined. A small number of families, such as *Tourist* I and *Stowaway* III, presented a particular problem—they contained over a 1000 intact elements, exceeding the number of sequence entries of ClustalW. Instead of haphazardly choosing about 500 elements (feasible for ClustalW), a more systematic approach was adopted. Each large group was divided into smaller groups, each of these was aligned, and an equal number of members from each branch of the alignment were included in the final, reduced set of sequences. This approach ensured that no linkage group was over-represented in the final alignment. Forty-eight profile HMMs were built.

A major drawback of using HMMER (or SAM) to analyze DNA sequences is that, since these packages were designed with protein analysis in mind, they do not take into account strandedness, searching only the top strand of a DNA sequence. Some MITEs are virtually palindromic so searching one or both strands will produce the same result; however, for most families a second profile had to be built using the reverse complement of the family alignment. In all, 94 profile HMMs were run on the 12 pseudomolecules and the results from the two profiles for each family were subsequently merged.

HMMER profile building includes an optional calibration step, which increases the search sensitivity. Searching with a calibrated profile produces the same hits with the same scores, but the *E*-values are somewhat decreased. Since the threshold for real hits was empirically determined for each TE family by human expert screening, the calibration step was deemed redundant.

The MULE set comprised 760 elements grouped into 40 families. MULEs are a class II TE superfamily designation that includes both autonomous elements that encode a transposase and non-autonomous elements that have suffered internal deletions. MULEs belong to the IS*256*/*Mutator* superfamily (Eisen *et al*., 1994). Most elements are characterized by long (about 200 bp) TIRs and a highly variable internal region. MULEs lacking TIRs (non-TIR MULEs) have been described in Arabidopsis (Yu *et al*., 2000), but have not been

found in rice (Le *et al*., 2000). In addition, many MULE elements have acquired stretches of host sequence, which can be either coding or non-coding. Thus, MULEs presented a challenge for profile building. Before performing the alignments, elements that lacked one or both TIRs were removed, and those with longer internal sequences were blasted against the NCBI nucleotide database to exclude sequences with similarities to host genes. The remaining sequences were handled as described for MITEs, resulting in 36 profile HMMs including seven reverse complement profiles. After the profiles were used to annotate the rice pseudomolecules, it became clear that there were numerous examples of single elements being annotated as multiple partial hits, as a consequence of MULE hypervariable internal regions that did not fit the model. To obtain a more realistic figure on the total number of MULEs, a simple heuristic was used to process the HMMER output: each instance of partial hits in the correct orientation (i.e. head optionally followed by internal region and tail), corresponding to the same MULE family and spanning less than 4 kb was counted as a single element.

At 1122 sequences, Ty3/*gypsy* was the second most numerous group in RTEdb. Since these elements can be more than 15 kb long, it was not feasible to run them with HMMER (it is possible to build profiles based on alignments longer than 1000 bp, but the time needed to run the search would increase dramatically). However, their number was reduced by first removing all elements shorter than 5 kb (i.e. partial elements), aligning the remaining elements, and removing those that seemed to be nearly identical to other elements. This procedure lowered the number of Ty3/*gypsy* entries in RTEdb to 23, reducing the final size of database size to 2 MB, from the initial 10.6 MB.

Table 1. Comparison of RepeatMasker and HMMER performance in the annotation of TEs

Feature	RepeatMasker	HMMER
Initial steps required	Compiling sequences	Compiling sequences Creating alignments Building profile HMMs
Speed	Slow with cross_match Faster with WU-BLAST, but still generally slower than HMMER for large families with short sequences	Alignments time-consuming Searches fast
Searching with long sequences	Yes	No
Searching both strands	Yes	No
Detecting diverged sequences	Less sensitive	More sensitive

CONCLUSIONS

Recent papers describing a draft of the rice genome sequence (Goff *et al*., 2002; Yu *et al*., 2002) or the finished sequence of one of the chromosomes (Feng *et al*., 2002; Sasaki *et al*., 2002; Rice Chromosome 10 Sequencing Consortium, 2003) used RepeatMasker for TE annotation. The source of TE sequences varied, but in the majority of cases RepBase (the default RepeatMasker database), TIGR RDB or a combination of the two was used. The results of our study, which show that using exclusively RepeatMasker is neither the most efficient nor the most sensitive approach, can serve as a reference for future genome projects.

Constructing and applying HMM approaches to large families of phylogenetically diverse TE repeats is a promising alternative approach to large-scale screening of genome sequences for such elements using simple alignment algorithms. Although constructing the required multiple sequence alignments to generate the models represents a significant upfront investment of time, once the models are constructed, they may be repeatedly reapplied efficiently to other genomes likely to carry similar repetitive elements. In the case of rice TE families, it is likely that genome sequence efforts for other, more highly repetitive monocot genomes such as maize will benefit greatly from this investment; however, the decision to apply profile methods like HMM in contrast to simpler alignment driven methods needs to address the tradeoff between profile and family size versus time complexity of simple alignments.

A summary of the comparison between RepeatMasker and HMMER is presented in Table 1. The performance of these methods needs to be investigated further, using different parameter settings. For instance, the default score cutoff of 200 was used for RepeatMasker, but this may not be optimal. Despite our benchmarking test, it is possible that the outcome would have been different if cross_match had been used instead of WU-BLAST. Finally, better quality multiple sequence alignments would have resulted in better quality profile HMMs and presumably more accurate results.

Using different methods to analyze different TE superfamilies raises one serious concern. If the assumption that profile HMMs can detect more degenerated elements is correct, using them generates higher numbers than using RepeatMasker. Since only class II elements were used to build profiles, this may skew the ratio between class I and class II elements, overestimating the relative contribution of class II. A unified method integrating the strengths but eliminating the flaws of presently available methods is needed. Obtaining the most accurate annotation would require combining an approach based on sequence information, such as an implementation of RepeatMasker using a profile HMM database, with *de novo* approaches and methods that look at TE-specific structural features.

ACKNOWLEDGEMENTS

We would like to thank Alexander Cosico, Victor Jun Ulat and Locedie Mansueto for technical assistance and advice and Maria Lalous for critical reading of the manuscript.

REFERENCES

Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Bao,Z. and Eddy,S.R. (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.*, **12**, 1269–1276.

Bedell,J.A., Korf,I. and Gish,W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**, 1040–1041.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Eddy,S.R. (2003) HMMER User's Guide. http://hmmer.wustl.edu

Eisen,J.A., Benito,M.I. and Walbot,V. (1994) Sequence similarity of putative transposases links the maize Mutator autonomous elements and a group of bacterial insertion sequences. *Nucleic Acids Res.*, **22**, 2634–2636.

Feng,Q., Zhang,Y., Hao,P., Wang,S., Fu,G., Huang,Y., Li,Y., Zhu,J., Liu,Y., Hu,X. *et al*. (2002) Sequence and analysis of rice chromosome 4. *Nature*, **420**, 316–320.

Goff,S.A., Ricke,D., Lan,T.H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* ssp. *japonica*). *Science*, **296**, 92–100.

Harushima,Y., Yano,M., Shomura,A., Sato,M., Shimano,T., Kuboki,Y., Yamamoto,T., Lin,S.Y., Antonio,B.A., Parco,A. *et al*. (1998) A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics*, **148**, 479–494.

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*. IEEE Computer Society Press, Los Alamitos, CA, Vol. 1, pp. 792–802.

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.

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.

Lazareva-Ulitsky,B. and Haussler,D. (1999) A probabilistic approach to consensus multiple alignment. *Pac. Symp. Biocomp.*, 150–161.

Le,Q.H., Wright,S., Yu,Z. and Bureau,T. (2000) Transposon diversity in *Arabidopsis thaliana. Proc. Natl Acad. Sci. USA*, **97**, 7376–7381.

Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, **30**, 4321–4328.

Mao,L., Wood,T.C., Yu,Y., Budiman,M.A., Tomkins,J., Woo,S., Sasinowski,M., Presting,G., Frisch,D., Goff,S., Dean,R.A. and Wing,R.A. (2000) Rice transposable elements: a survey of 73 000 sequence-tagged-connectors. *Genome Res.*, **10**, 982–990.

McCarthy,E.M., Liu,J., Lizhi,G. and McDonald,J.F. (2002) Long terminal repeat retrotransposons of *Oryza sativa. Genome Biol.*, **3**, research0053.1–0053.11.

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. and Lipman,D. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Pedersen,J.S. and Hein,J. (2003) Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, **19**, 219–227.

Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. *Science*, **300**, 1566–1569.

Saji,S., Umehara,Y., Antonio,B.A., Tanoue,H., Yamane,H., Baba,T., Aoki,H., Ishige,N., Wu,J., Koike,K., Matsumoto,T. and Sasaki,T. (2001) A physical map with yeast artificial chromosome (YAC) clones covering 63% of the 12 rice chromosomes. *Genome*, **44**, 32–37.

Sasaki,T., Matsumoto,T., Yamamoto,K., Sakata,K., Baba,T., Katayose,Y., Wu,J., Niimura,Y., Cheng,Z., Nagamura,Y. *et al*. (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.

Sinha,S., van Nimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**, i292–i301.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Turcotte,K., Srinivasan,S. and Bureau,T. (2001) Survey of transposable elements from rice genomic sequences. *Plant J.*, **25**, 169–179.

Yamamoto,K. and Sasaki,T. (1997) Large-scale EST sequencing in rice. *Plant Mol. Biol.*, **35**, 135–144.

Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.

Yu,Z., Wright,S.I. and Bureau,T.E. (2000) Mutator-like elements in *Arabidopsis thaliana*: structure, diversity and evolution. *Genetics*, **156**, 2019–2031.