Heterologous approach in the search for (candidate) genes

J. Žlůvová

Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic

ABSTRACT

To answer many questions of evolutionary developmental biology, it is necessary to identify genes which could be responsible for the studied traits. In this review, I have summarised the methods that are currently used to identify these genes. Mainly the PCR-based methods of gene cloning are discussed here, with the special attention given to the comparison of primer design strategies and PCR amplification strategies. In addition to the review of methods used by other authors, an overview of our experience with PCR-based gene cloning in dioecious model plant species *Silene latifolia* is included.

Keywords: evolutionary developmental biology; degenerate primer design; PCR

Why are there bilaterally symmetric flowers in pansy (*Viola*) instead of radially symmetric ones as in poppy *(Papaver)* and why are there lobed leaves in oak (*Quercus*) instead of entire ones as in *Tillandsia*? To answer these and many other questions of evolutionary developmental biology, it is very difficult to use "classical" approaches like mutants screens (as many morphological novelties are caused by a loss of function, and because of the genetic redundancy which is a relatively frequent phenomenon), positional cloning (only few data are available in the non-model species), or subtraction cloning (too many genes will be obtained from subtractions performed between two distantly related species). The straightest method is based on cDNA or genomic library sequencing, followed by an analysis of sequence data and comparative expression profiling to identify genes that are candidates for a causal role in evolutionary divergence. This approach was performed for a small group of species by the Floral Genome Project (Floral Genome Project Research Group 2002). However, it is still extremely costly and time-consuming.

A direct search for candidate genes followed by their subsequent analysis seems to be a relatively cheap and fast alternative. Using this strategy, e.g. the signalling pathway leading to the unifoliate growth form in *Streptocarpus* was found (Harrison et al. 2005). The first, and very often also the most difficult, step is obtaining the candidate genes; hence, this review has focused on different methods used in the search for candidate genes.

To search directly for candidate genes, it is possible to screen cDNA (or genomic) library using probes coming from other species. However, some difficulties can appear. Appropriate libraries are not available in many non-model organisms, and thus it is often necessary to prepare a new library. Moreover, a failure in the screening can appear if the desired gene is not highly conserved. Additionally, when cDNA libraries are screened, a serious problem can be caused by the fact that some transcripts are present at exceedingly low levels in the tissues in which they act. For example, the gene *TEOSINTE-BRANCHED1* (a gene expected to play a crucial role in the evolution of maize from its wild progenitors, Doebley et al. 1997) has not been identified in the public maize database of 100 000 EST's (Baum et al. 2002).

Most of these difficulties can be overcome by the PCR-based search for candidate genes: it is not necessary to construct libraries, the probability that conserved parts of 20–30 nucleotides appropriate for primer design will be found is much higher than the probability that a considerable part of a gene

Supported by the Czech Science Foundation, Project No. 204/05/P505, and by the Institutional Research Plan, Project No. AVOZ50040507.

This contribution was presented at the 4th Conference on Methods in Experimental Biology held in 2006 in the Horizont Hotel in Šumava Mts., Czech Republic.

is conserved enough to be found by hybridisation, and PCR is enough sensitive to amplify transcripts of genes expressed at extremely low level.

PROTOCOLS AND TROUBLESHOOTING

Primer design strategies

Primer design can be still very difficult because of codon degeneracy and the additional degeneracy needed to represent multiple codons at a position in the alignment. Basically, three primer design strategies were employed previously (reviewed in Rose et al. 1998). One strategy is to design consensus primers across highly conserved regions. However, primer-to-template mismatches in distantly related species limit this strategy mostly to closely related species. The second strategy is to synthesize a pool of degenerate primers containing most or all of the possible nucleotide sequences implicit in a multiple alignment. A serious problem is that as the degeneracy increases, the concentration of any single primer drops. As a result, the number of primer molecules in the PCR that can prime synthesis during the amplification cycles drops, and these primers are used up early in the reaction. This strategy is used in the primer design programs PriFi (Fredslund et al. 2005) and Primaclade (Gadberry et al. 2005).

The strategy called COnsensus DEgenerate Hybrid Oligonucleotide Primers (CODEHOP) (Rose et al. 1998, Rose et al. 2003a) seems to overcome disadvantages of both degenerate and consensus primer design methods. Primers consist of a relatively short (11–12 nucleotides) 3' degenerate core and a long (more than 23 nucleotides) 5' non-degenerate consensus clamp. Primers are designed according to conserved amino acid sequences. The 3' degenerate core contains all of the possible nucleotide sequences whereas the 5' non-degenerate consensus clamp contains the most probable nucleotides within each codon. Reducing the length of the 3' core to a minimum decreases the total number of individual primers in the degenerate primer pool. Hybridization of the 3' degenerate core with the target template is stabilized by the 5' non-degenerate consensus clamp, which allows higher annealing temperatures without increasing the degeneracy of the pool. A program for automatically predicting optimal primers that embody the CODEHOP strategy, implemented at the World Wide Web, is available at: http://bioinformatics.weizmann.ac.il/blocks/codehop.html.

However, an ideal, generally applicable primer design strategy does not exist. The situation is complicated by the fact that genes contain not only sequences conserved on the amino acid level (i.e. conserved protein domains) but also sequences conserved on the DNA level (e.g. microRNA binding sites). Limits of both CODEHOP and cDNA-based strategies can be illustrated by an example from our laboratory work. Two different primer design strategies were applied on problems in which the target sequences for amplification were not known but could be predicted from multiply aligned protein and cDNA sequences. The goal was to clone eleven different genes involved in floral meristem formation from *Silene latifolia*. When CODEHOP strategy was employed, four genes out of nine tested were cloned. When degenerate primers designed according to conserved cDNA sequences were used, six genes out of nine tested were cloned. Two genes were identified exclusively using classical degenerate primers, but not using CODEHOP primers. Three genes were not amplified by either of these strategies, although eight different primer pairs distributed along the full gene sequences of both CODEHOP and classical degenerate primers were used for each gene. It is possible to hypothesise that the conservancy of these genes between different plant species is relatively low.

As we have subsequently performed 5' and 3' RACE-PCR to obtain complete sequences of the amplified fragments, it is possible to conclude that the differences in the PCR success between CODEHOP and classical degenerate primers were caused by the fact that the conserved amino acid blocks used for the CODEHOP primer design were not sufficiently conserved in *S. latifolia* genome to enable PCR amplification. Morant et al. (2002) came to similar conclusions when amplifying *CYP98* in the *Solanaceae* family. As the dissimilarities in the nucleotide sequences are likely to be local, successful amplification should be obtained by using multiple primers distributed along the full sequence.

PCR template

In principle, both genomic DNA and cDNA can be used as a PCR template to find the desired genes. A big advantage of the genomic DNA as a template is that a sample coming from any tissue contains complete genetic information. Thus, it is not necessary to isolate DNA from tissues

where the desired gene is expressed. Genomic DNA as a template has also one big disadvantage – it contains complete genetic information. Because degenerate primers are used, the probability of artifactual amplification of "junk DNA" is high because of the dominance of primers in the pool, which do not participate in amplification of the targeted gene but are available to prime non-specific synthesis. As a result, after the PCR it leads to an extremely high amount of bands on a gel. On the other hand, although it is necessary to extract RNA from the tissue where the desired gene is expressed, the enormous advantage is that it (theoretically) does not contain any "junk DNA" and the RT-PCR with degenerate primers relatively often leads to a single band on a gel.

PCR amplification strategies

As primer positions are fixed to the conserved regions, it is often impossible to obtain primers of optimal PCR performance. To empirically determine optimal annealing and amplification conditions for the pool of primers, thermal gradient PCR amplification can be used (e.g. Rose et al. 2003b). PCR amplification can be performed using not only classical but also touchdown approaches (e.g. Rose et al. 1998). In the case of highly conserved genes and extremely well-designed primers, the temperature can range from 60°C to 50°C (e.g. Zluvova et al. 2005), but PCR gives very rarely any product under this condition. A solution is to perform touchdown PCR from 60°C to 20°C followed by 20 to 30 PCR cycles under 20°C annealing temperature. In this case, it is necessary to add DNA polymerase to the PCR mix twice – before the beginning of PCR and after the touchdown cycles. We have used both more and less stringent PCR conditions to identify *S. latifolia* orthologue of the gene *SHOOTMERISTEMLESS* (Zluvova et al. 2006). When more strict PCR conditions were used [3 min at 94°C followed by 10 cycles of 30 s at 94°C, 1 min at 60°C (minus 1°C at each cycle) and 2 min at 72°C, followed by 25 cycles of 30 s at 94°C, 1 min at 50°C, 2 min at 72°C, followed by the final extension of 7 min at 72°C], PCR did not give any product of the expected length. When we employed less strict PCR conditions [3 min at 94°C followed by 25 cycles of 30 s at 94°C, 1 min at 60°C (minus 1.6°C at each cycle) and 1 min at 72°C, followed by addition of Taq polymerase, followed by 15 cycles of 15 s at 94°C, 1 min at 20°C and 1 min at 72°C, followed by the final extension of 7 min at 72°C], we obtained PCR products of expected length, and subsequent analyses confirmed that the obtained sequence is orthologous to the gene *SHOOTMERISTEMLESS*.

Although we have very good experience with the above-described PCR strategy, it gives relatively often multiple bands or smears. We have partially solved this problem by a combination of thermal gradient PCR and 10°C temperature decrease (in total) during the touchdown PCR cycles.

Still, the reaction can give a smear or no PCR product. A possible solution is to perform nested or semi-nested PCR. Because the primers are degenerated, this method can lead to non-specific bands or smears, and optimisation of both the first and second PCR cycles often solves this problem. A disadvantage of this double optimisation is an enormous increase of PCR samples to be analysed.

When using a combination of both degenerate and specific primers (the specific primer matches for example microRNA binding site), it can be difficult to find any overlap between annealing temperatures of both primers. In this case, it is good to perform PCR in two steps. The first step of 20–30 cycles is done with the specific primer only and the annealing temperature is high. This step enables linear extension from the specific primer. The second step is done with both the specific and degenerate primers, under a low annealing temperature. This step enables extension from the degenerate primer and also amplification of the desired gene. We have employed this PCR strategy to amplify *S. latifolia* orthologue of the gene *CUP-SHAPED COTYLEDONS* containing microRNA binding site conserved on the DNA level and a NAC domain conserved on the protein level (Zluvova at al. 2006). The first step performed with the specific primer was as follows: initial denaturation (3 min at 94°C) followed by 18 cycles of 20 s at 94°C, 1 min at 65°C (minus 0.3°C at each cycle) and 1 min at 72°C. The second step with both primers was as follows: 18 cycles of 15 s at 94°C, 1 min at 55°C (minus 0.5°C at each cycle) and 1 min at 72°C, followed by 18 cycles of 15 s at 94°C, 1 min at 55°C and 1 min at 72°C followed by the final extension of 7 min at 72°C.

Hypothesis testing

The obtained PCR products can be directly sequenced only when the desired gene is not a member of a gene family. Otherwise, it is better to

clone the PCR product and to sequence several colonies. After the sequencing, orthology of the cloned sequence has to be verified. As the BLAST or FASTA searches are usually not sufficient to unambiguously confirm the orthology, construction of phylogenetic trees is usually used (reviewed in Michu 2007).

Because the initial working hypothesis was that the cloned orthologue diverged from its copies studied in model plants, and that it plays a different role in our non-model species, the phylogenetically identified orthologue should be further characterised with respect to its expression pattern. The easiest and most commonly used method is RT-PCR, which provides an overall view on the expression pattern. Much more information can be obtained from *in situ* hybridisation experiments, which detect mRNA on tissue sections, so that the location of the target molecule can be seen in individual cells.

Perspectives

The search for orthologous sequences is not only the first step in the evolutionary-developmental biology, but can be also used in other disciplines. The molecular genetic markers, which can be transferred between species, can be used to study syntenic relationships in the field of comparative genomics. A broad application of the method is also in agricultural breeding programs, because such markers can be used to optimize the exploitation of genetic map resources. The method can be also used as a source of molecular markers in mapping and breeding experiments (e.g. marker-assisted breeding), which are often marker-limited due to the lack of DNA sequence information from the species in question.

REFERENCES

Baum D.A., Doebley J., Irish V.F., Kramer E.M. (2002): Response: Missing links: the genetic architecture of flower and floral diversification. Trends Plant Sci., *7:* 31–34.

Doebley J., Stec A., Hubbard L. (1997): The evolution of apical dominance in maize. Nature, *386:* 485–488.

Floral Genome Project Research Group (2002): Missing links: the genetic architecture of flower and floral diversification. Trends Plant Sci., *7:* 22–31.

Fredslund J., Schauser L., Madsen L.H., Sandal N., Stougaard J. (2005): PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. Nucleic Acids Res., *33:* W516–W520.

Gadberry M.D., Malcomber S.T., Doust A.N., Kellogg E.A. (2005): Primaclade – a flexible tool to find conserved PCR primers across multiple species. Bioinformatics, *21:* 1263–1264.

Harrison J., Moller M., Langdale J., Cronk Q., Hudson A. (2005): The role of *KNOX* genes in the evolution of morphological novelty in *Streptocarpus*. Plant Cell, *17:* 430–443.

Michu E. (2007): A short guide to phylogeny reconstruction. Plant Soil Environ. (submitted)

Morant M., Hehn A., Werck-Reichhart D. (2002): Conservation and diversity of gene families explored using the CODEHOP strategy in higher plants. BMC Plant Biol., *2:* 7.

Rose T.M., Henikoff J.G., Henikoff S. (2003a): CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Res., *31:* 3763–3766.

Rose T.M., Ryan J.T., Schultz E.R., Raden B.W., Tsai C.C. (2003b): Analysis of 4.3 Kb of the divergent locus-B of macaque retroperitoneal fibromatosis-associated herpesvirus (RFHV) reveals close similarity to Kaposi's sarcoma-associated herpesvirus (KSHV) in gene sequence and genome organization. J. Virol., *77:* 5084–5097.

Rose T.M., Schultz E.R., Henikoff J.G., Pietrokovski S., McCallum C.M., Henikoff S. (1998): Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res., *26:* 1628–1635.

Zluvova J., Lengerova M., Markova M., Hobza R., Nicolas M., Vyskot B., Charlesworth D., Negrutiu I., Janousek B. (2005): The inter-specific hybrid *Silene latifolia* × *S. viscosa* reveals early events of sex chromosome evolution. Evol. Dev., *7:* 327–336.

Zluvova J., Nicolas M., Berger A., Negrutiu I., Moneger F. (2006): Premature arrest of the male meristem precedes sexual dimorphism in the dioecious plant *Silene latifolia*. Proc. Natl. Acad. Sci. USA, *103:* 18854–18859.

Received on February 27, 2007

Corresponding author:

Mgr. Jitka Žlůvová, Ph.D., Akademie věd České republiky, Biofyzikální ústav, v. v. i., Královopolská 135, 612 65 Brno, Česká Republika

phone: + 420 541 517 247, fax: + 420 541 240 500, e-mail: jitka@ibp.cz