Available online at www.sciencedirect.com
ScienceDirect
Journal of Biotechnology 130 (2007) 1–10
Journal of BIOTECHNOLOGY
www.elsevier.com/locate/jbiotec

Sequence-based factors influencing the expression of heterologous genes in the yeast *Pichia pastoris*—A comparative view on 79 human genes

Mewes Boettner[a,1], Christina Steffens[a,1], Christian von Mering[b,2], Peer Bork[b], Ulf Stahl[a], Christine Lang[a,*]

[a] *Berlin University of Technology, Institute for Biotechnology, Department of Microbiology and Genetics, Gustav-Meyer-Allee 25, D-13355 Berlin, Germany*
[b] *EMBL Heidelberg, Meyerhofstrasse 1, D-69117 Heidelberg, Germany*

Received 27 June 2006; received in revised form 16 February 2007; accepted 22 February 2007

Abstract

High yield expression of heterologous proteins is usually a matter of "trial and error". In the search of parameters with a major impact on expression, we have applied a comparative analysis to 79 different human cDNAs expressed in *Pichia pastoris*. The cDNAs were cloned in an expression vector for intracellular expression and recombinant protein expression was monitored in a standardized procedure and classified with respect to the expression level. Of all sequence-based parameters with a possible influence on the expression level, more than 10 were analysed. Three of those factors proved to have a statistically significant association with the expression level. Low abundance of AT-rich regions in the cDNA associates with a high expression level. A comparatively high isoelectric point of the recombinant protein associates with failure of expression and, finally, the occurrence of a protein homologue in yeast is associated with detectable protein expression. Interestingly, some often discussed factors like codon usage or GC content did not show a significant impact on protein yield.

These results could provide a basis for a knowledge-oriented optimisation of gene sequences both to increase protein yields and to help target selection and the design of high-throughput expression approaches.

Keywords: *Pichia pastoris*; Heterologous protein expression; Sequence-based factors

1. Introduction

The heterologous expression of proteins is very often a substantial part of biochemical studies in all fields and it is none the less vital for numerous biotechnological applications (Garber, 2001). Despite increasing experience in protein expression, it is still a time and labour consuming "trial and error" approach to find the right combination of host and protein and to improve the protein yield when expression is finally successful (Stevens, 2000; Yokoyama, 2003). Very often, this part of a project takes the major part of resources. Up to now, there are no reliable criteria for predicting the success of expression of a particular protein in a given host. We were able to approach this well-known bottleneck by a comparative view on a large set of human cDNAs that were identically cloned and expressed, with respect to their expression level in the widely used yeast host *Pichia pastoris*.

In the framework of the Protein Structure Factory – a German structural genomics initiative – (Heinemann et al., 2000) a collection of yeast expression clones harbouring human cDNAs was established (Holz et al., 2003).

The cDNAs were expressed as fusion proteins with an N-terminal His_6- and a C-terminal StrepII-tag for intracellular expression in *Saccharomyces cerevisiae* and *P. pastoris* (Boettner et al., 2002; Holz et al., 2003). The *P. pastoris* clones were assigned to one of four categories of expression level (none, low, medium, and high expression) (Boettner et al., 2002).

This collection of expression clones screened and evaluated using standardized procedures holds identical strains differing only in the respective cDNA insert. This allows for a view on sequence-based parameters that might influence the expression level in the host. Seventy nine clones carrying different

* Corresponding author. Tel.: +49 30 31472751; fax: +49 30 31472922.
E-mail address: christine.lang@tu-berlin.de (C. Lang).
[1] Present address: OrganoBalance GmbH, Gustav-Meyer-Allee 25, D-13355 Berlin, Germany.
[2] Present address: Institute of Molecular Biology, University of Zurich, Winterthurerstrasse 190, CH-8057, Zurich, Switzerland.

0168-1656/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2007.02.019

human cDNAs were analysed in detail. Parameters previously described to affect gene expression and more general sequence-based features were analysed and related to the expression level. These included AT-rich regions conferring premature transcription termination in yeast (Zhao et al., 1999), GC-rich regions, the overall nucleotide composition, the general codon usage (Wright, 1990), the codon adaptation to yeast (Bennetzen and Hall, 1982), general protein features and signals known to mediate protein degradation in vivo, e.g. PEST motifs (Rechsteiner and Rogers, 1996). The potential influence of similarity to yeast proteins (*S. cerevisiae*) was also investigated.

Our data reveal for the first time that it is indeed possible to identify protein-associated or sequence-based factors by looking for common features of sequences leading to comparable expression results.

These results will eventually provide information regarding the choice of *P. pastoris* as a suitable host system and the knowledge-based optimisation of gene sequences to increase the protein yield. This approach might prove a valuable bioinformatic system to save time and costs for expression trials.

2. Materials and methods

2.1. Selection of targets

The selection of cDNAs was done as previously described (Holz et al., 2003). Briefly, criteria for selection were the following: predicted proteins smaller than 500 amino acids, no structure deposited in the Protein Data Bank, no transmembrane domains, no coiled-coil regions, and no compositional bias. After removing redundant sequences from the preselected cDNAs, an EST database containing all clones from the I.M.A.G.E. consortium (Lennon et al., 1996) was screened using BLAST to identify full-length cDNA clones available.

2.2. Cloning of cDNAs and screening of transformants

P. pastoris expression clones were constructed as described previously (Boettner et al., 2002). Briefly, cDNAs were amplified using ProofStart DNA Polymerase (Qiagen, Germany). Primers for cDNA amplification contained restriction sites for *Bam*HI or *Bgl*II (forward primer; restriction site depends on the cDNA sequence) and *Not*I, respectively (reverse primer). After restriction, inserts were ligated in frame into pPICHS (Boettner et al., 2002). Proteins are expressed as fusion proteins with an N-terminal His_6 and a C-terminal StrepII tag.

Screenings for protein expression were performed in a 2 ml scale using 24-deep well plates (Whatman, UK) as described (Boettner et al., 2002). Briefly, two individual transformants for each cDNA were precultivated in WM9 media for three days, shifted to fresh WM9 media without carbon source and expression was induced by adding 1% (v/v) methanol and 0.1% (w/v) glucose (final concentrations).

Expression was monitored after 24 h by Western blotting. Estimation of expression was done by visual comparison of the signal strengths with a standard clone expressing human Med7 (GenBank Accession: AAC52115). Detection of expression was done as a first step using penta-His antibody (Qiagen, Germany). Membranes were stripped and reprobed using StrepTactin-peroxidase conjugate (IBA, Germany). Usually both detection methods resulted in the same classification.

The clones were classified according to the relative signal strength into four different classes: no detectable expression (marked as: −), low expression (+), medium expression (++) and high expression (+++). The standard expression clone used for comparison gave a relative expression yield of ++ and therefore allowed a reliable categorization of the other expression levels (see Table 1 for protein identifiers and expression level).

2.3. Sequence analysis

2.3.1. Quantification of AT- and GC-rich regions

Plots of nucleotide composition along the sequences were generated by using the program freak (EMBOSS suite) (Rice et al., 2000) via the worldwide web access of Institute Pasteur, France (http://www.pasteur.fr/). Residue letters were set to AT or GC, respectively, stepping value was 1 and window size was set to 30. To get a quantification of, for example, AT-rich regions rather than average values of the whole sequence, the resulting tables were processed as follows: nucleotide positions that were assigned a value of 0.6 or higher by freak were regarded as part of an AT-rich region. Only those values assigned to these positions were added along the whole cDNA sequence. The resulting value was taken as a combined measurement of the length, the AT fraction of AT-rich regions and the fraction of those within the respective sequence. To compare the distribution of the sum of AT-rich regions with the distribution of the overall AT fraction of the cDNAs, the latter fraction was calculated as well. GC-rich regions were determined accordingly.

2.3.2. Codon usage as compared to S. cerevisiae

Codon usage was measured using the program CodonW (available via http://www.molbiol.ox.ac.uk/cu/) by John Peden. Calculated parameters were the number of effective codons (Wright, 1990) and the codon adaptation index (Sharp and Li, 1987). Due to the lack of genomic information and expression data of *P. pastoris*, the latter was measured towards a set of highly expressed genes in the yeast *S. cerevisiae* (Sharp and Cowe, 1991).

2.3.3. Distribution of rare codon as used for S. cerevisae

Codons were regarded as rare codons in yeast according to Zhang et al. (1991). For *S. cerevisiae* these are AGG (Arg); CGA (Arg); CGG (Arg); CGC (Arg); CCG (Pro); CUC (Leu); GCG (Ala); UCG (Ser). We calculated the frequency of these codons per coding sequence as well as the absolute occurrence per cDNA.

2.3.4. General protein features

Calculated protein features were the isoelectric point (Ribeiro and Sillero, 1991), the general average hydrophobicity (GRAVY) score, which is the arithmetic mean of the sum of the hydrophobic indices of each amino acid (Kyte and Doolittle,