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.
© 2007 Elsevier B.V. All rights reserved.

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,

Table 1
Expression levels and GenBank identification numbers of target proteins

	Occurence of homologous protein in yeast	p*I*	Peak area of AT-rich regions
No detectable expression			
AAA20587		8.9	138.0
AAA35648		10.2	92.9
AAA36597	+	9.2	11.7
AAA61187		10.0	6.1
AAA63269		10.4	74.5
AAA70088		5.1	217.9
AAA74903		7.8	30.4
AAA93231	+	8.5	124.0
AAB24206		7.8	7.9
AAB25225		7.7	200.8
AAB38529		5.9	10.5
AAB64192	+	8.5	57.5
AAB81453		6.4	45.5
AAB88175		6.2	466.5
AAC18356		8.9	223.0
AAC27445		6.6	3.7
AAC34987		8.3	161.4
AAC35550		8.8	73.1
AAC39912	+	6.3	95.8
AAC51284		6.3	113.3
AAD03265		6.4	0.0
AAD08720	+	8.6	0.0
AAD11629		7.1	0.0
AAD16169		6.2	244.2
AAD20972		9.5	29.0
AAD21526		6.0	105.2
AAD27769		4.4	29.5
*AAD34095**		6.1	86.6
AAD34115		5.8	58.7
AAD34133		6.4	8.5
AAD44492		7.9	23.2
AAD49967	+	1.1	98.2
AAF03512		6.2	0.6
AAF15100		6.1	235.3
BAA03400	+	8.9	66.4
BAA05118		8.9	61.0
BAA11485		5.3	1.2
BAA33391	+	6.2	15.7
CAA22906		5.0	191.3
CAA51827		8.7	2.4
CAA58535		11.4	15.2
CAA65339		8.9	31.7
CAB52345		7.8	55.8
CAB56506		9.9	91.4
Low expression level			
AAA60286	+	9.8	65.6
AAA87395	+	5.8	102.9
AAB00114	+	5.8	0.0
AAC39715		7.2	36.4
AAD02685	+	6.6	3.6
AAD09623	+	7.9	94.2
AAD20048		6.1	65.6
AAD44363		6.3	190.9
AAD44489		6.0	9.1
AAD51801		5.6	200.8
AAF03537		5.9	24.8
AAF14857	+	6.6	28.5
AAF14877		5.0	207.6
BAA08392	+	5.6	309.0
BAA13402		9.0	0.0
Medium expression level			
AAB81205		5.2	297.4
AAC41945	+	5.8	62.7
*AAC52115**	+	6.2	270.0
AAC62536	+	7.7	110.8
AAD27741		7.2	94.0
AAD34115	+	9.1	12.8
AAF13149		6.3	123.1
BAA09317		5.7	14.7
BAA12872		5.9	44.5
CAA34200		9.4	0.0
CAA34379	+	9.1	40.9
CAA34890	+	5.9	71.9
CAA37376	+	6.8	100.4
CAB56175		5.8	28.0
High expression level			
AAB96936	+	7.8	0.6
AAC27445		5.9	4.2
AAC83329		4.6	0.0
*AAD25021**		6.4	0.0
AAD27777	+	6	0.0
AAF00499	+	5.4	64.6

The expression of cDNAs marked with an '*' has been published before in Boettner et al. (2002).

1982), the aromaticity, which is the frequency of aromatic amino acids, and the protein length.

2.3.5. *Protein degradation signals*

PEST motifs (Rogers et al., 1986) were determined using PESTfind (Rechsteiner and Rogers, 1996) (http://www.at.embnet.org/embnet/tools/bio/PESTfind/). Window size was set to 10. As output scores of the program vary from −50 to 50 for each PEST-like motif found, the scores were converted into the positive range by adding 50 to allow easier processing. A score of zero was assigned to cDNAs giving no PESTfind hit.

Lysine residues were taken into account as potential sites of ubiquitination and subsequent degradation (reviewed by Weissman, 2001). The surface probability of each lysine residue was taken into account to estimate the accessibility to the ubiquitination machinery. The surface probability was calculated using the Emini-Index (Emini et al., 1985) by the program DNAStar (DNASTAR Inc., WI). Lysine residues were assigned to have a high surface probability when the Emini-Index for the respective position was 1 or higher (Emini et al., 1985).

2.3.6. *Similarity of targets to yeast proteins*

All 79 target proteins were searched against the full proteome of *S. cerevisiae* (downloaded from the 'SPproteomes' database (Brooksbank et al., 2003)). The search was done using the Smith–Waterman algorithm, with low complexity filtering disabled and gap parameters set to −11 for opening, and −1 for extension. The presence of a true yeast homolog was assumed when the best-scoring alignment extended over at least 100 amino acid residues and scored at least 60 bits.

2.3.7. Statistical evaluation

Due to the fact that the data are not normally distributed and the ordered categorisation – except for the data regarding the occurrence of a homologue in yeast – a nonparametric Kruskal–Wallis test (Kruskal and Wallis, 1952) was applied for evaluating statistically significant differences between the categories of expression level. The significance of the distribution of yeast homologues was evaluated using the chi-square test.

3. Results

3.1. Distribution of expression level

The cDNAs were cloned, the respective *P. pastoris* transformants were tested for protein expression and the expression level was quantified as described before (Boettner and Lang, 2004). Sequences were considered for detailed analysis when the size of the amplified cDNA was as expected and the two yeast transformants screened per cDNA behaved uniformly. Detection of the N-terminus – using anti-penta-His antibody (Qiagen, Germany) – as well as of the C-terminus – using StrepTactin-peroxidase conjugate (IBA, Germany) – was performed in all cases. All of the proteins detected were positive for both tags and usually both the detections resulted in the same classification of expression level. Therefore, 79 cDNA sequences were analysed further. The screening resulted in 44 cDNAs showing no expression, 15 showing low, 14 showing medium, and six showing high expression level (see Table 1).

3.2. AT- and GC-rich regions

Our analysis revealed that a high expression level is associated with a low amount of AT-clusters within the coding sequence ($P = 0.036$; Fig. 1a). The other categories of expression level (none, low, medium) do not differ significantly in the amount of AT-rich regions.

There is no significant association with the amount of GC-rich regions with either category of expression level ($P = 0.172$).

This is true both for the absolute value calculated for the whole sequence and for the value relative to the length of the sequence in nucleotides ($P = 0.037$; Fig. 1b).

There is no significant association of the amount of GC-rich regions with either category of expression level ($P = 0.172$).

The association of AT-rich regions with a high expression level is not reflected in the overall GC-fraction ($P = 0.749$; Fig. 1c), or the GC-content at the silent third codon position ($P = 0.806$; Fig. 1d). This shows that the observed association between a low value of AT-clusters and a high expression level is neither a side effect of the overall AT-content nor of base preferences at silent third codon positions.

3.3. Codon usage

To measure codon bias and codon adaptation to yeast we computed the number of effective codons (Nc) (Wright, 1990) and the codon adaptation index (CAI) (Sharp and Li, 1987). The latter was measured towards a set of highly expressed *S. cerevisiae* genes (Sharp and Cowe, 1991). A recent consideration of genome-wide expression data sets revealed the CAI derived from the data of Sharp and Cowe (1991) to be fairly insensitive to this extension to genome-wide data sets (Jansen et al., 2003). These two indices were chosen among various possibilities to measure codon usage or bias because – in contrast to other existing indices – they give equivalent values independent of the length of the respective sequences (Comeron and Aguade, 1998). It has previously been shown that preferred codons are very similar to those of *S. cerevisiae*, at least in the few *P. pastoris* genes known to be highly expressed (Sinclair and Choy, 2002). Therefore, measuring codon adaptation towards a *S. cerevisiae* set of highly expressed genes seems a valid approach. The results are shown as box plots in Fig. 2 and were evaluated using the Kruskal–Wallis test (Kruskal and Wallis, 1952).

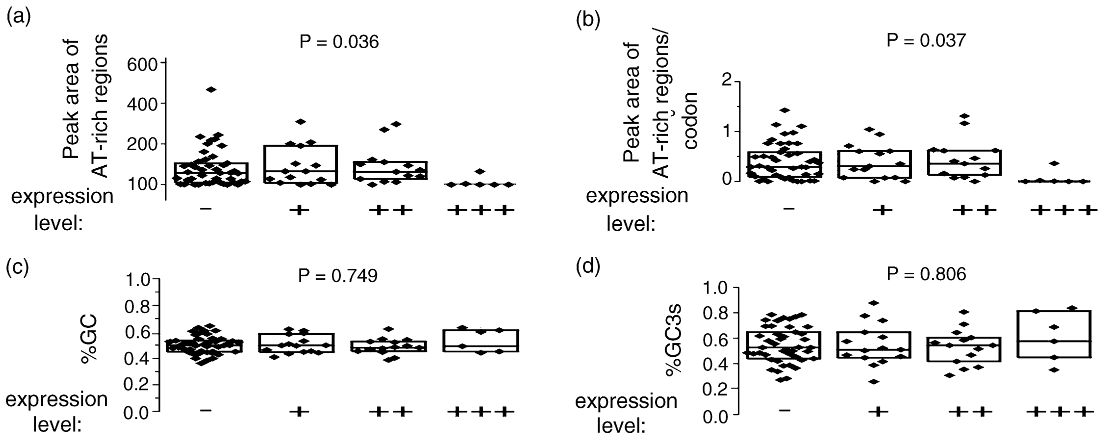

Fig. 1. Distribution of AT-rich clusters, GC-content and GC3s within the categories of expression. Plots of nucleotide composition along the sequences were generated by using the program freak (EMBOSS suite). To get a quantification of AT-rich regions rather than average values of the whole sequence, nucleotide positions that were assigned a value of 0.6 or higher by freak were regarded as part of an AT-rich region. Those values assigned to these positions were added along the whole cDNA sequence. The resulting value represents a peak area of an AT-rich region within the sequence. The peak area of AT-rich clusters is associated with a high expression level. This is true for the peak area of AT-rich regions ($P = 0.036$) (a) as well as for the peak area of AT-rich regions per number of codons within the respective cDNA ($P = 0.037$) (b). This association is not reflected by the distribution of the overall GC-content (c) nor the GC-content at the silent third codon position (GC3s) (d).

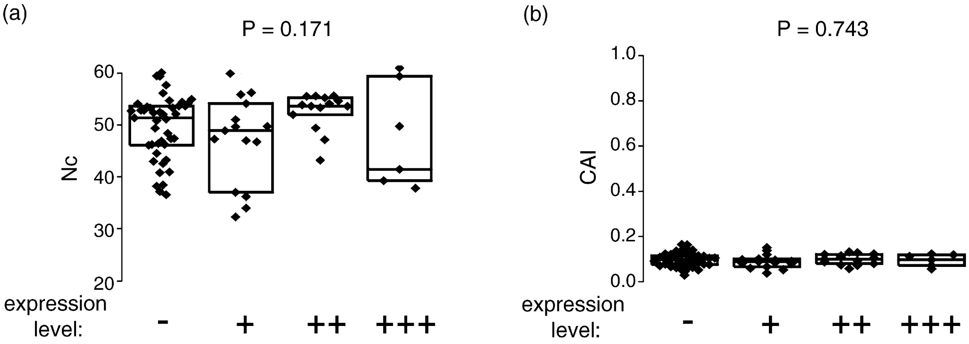

Fig. 2. Distribution of codon bias and codon adaptation to yeast within the categories of expression. Calculated parameters were the number of effective codons (Nc) – a measure for the codon bias within a given sequence – and the codon adaptation index (CAI). Due to the lack of genomic information and expression data of *P. pastoris*, the latter was measured towards a set of highly expressed genes in the yeast *Saccharomyces cerevisiae*. Neither the number of effective codons (a) nor the codon adaptation to yeast (b) are associated to a level of expression for the range of those indices observed.

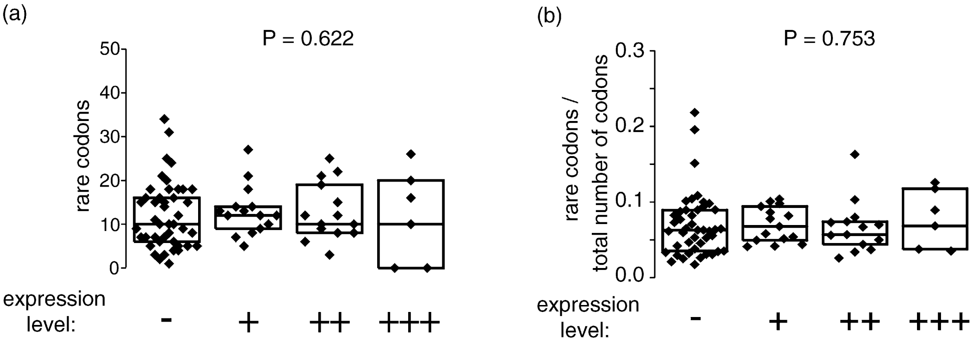

Fig. 3. Distribution of codons rare in yeast within the categories of expression. Codons were regarded as rare codons in yeast according to Zhang et al. (1991). Shown is the amount of rare codons per coding sequence (a) as well as the frequency of rare codons (b). None of those measures is associated to a level of expression.

An association between codon usage and expression level is not detectable. Neither codon bias ($P = 0.171$; Fig. 2a) nor a codon adaptation towards codon usage of *S. cerevisiae* ($P = 0.743$; Fig. 2b) is associated with one specific group of expression level.

Fig. 3a and b depicts the occurrence of non-preferred codons in yeast as well as their frequency per cDNA. No significant differences in the amount of rare codons per coding sequence are detectable between the groups of expression level ($P = 0.622$; Fig. 3a). This is also observed when the frequency of rare codons per total number of codons is analysed ($P = 0.753$; Fig. 3b). Thus, taking into account the cDNAs tested in this study, the codon usage as well as the occurrence of unfavourable codons is not a major factor influencing expression yield.

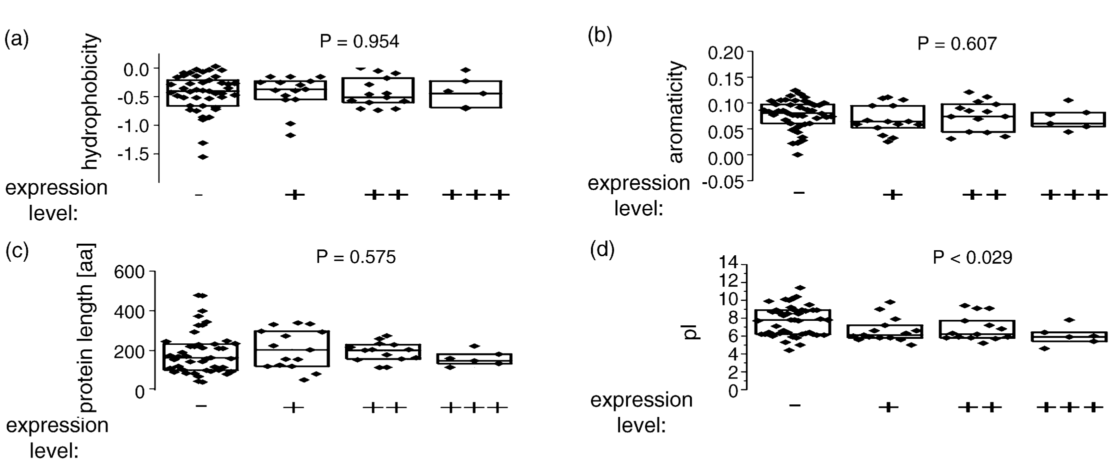

Fig. 4. Distribution of general protein features, calculated protein features were the isoelectric point (p*I*), the general average hydrophobicity (GRAVY) score, which is the arithmetic mean of the sum of the hydrophobic indices of each amino acid, the aromaticity, which is the frequency of aromatic amino acids, and the protein length. Neither the hydrophobicity (a), the aromaticity (b) nor the length (c) of the protein is associated with the expression level. A significant association ($P = 0.029$) has been observed between a high isoelectric point (p*I*) and a non detectable expression (d).

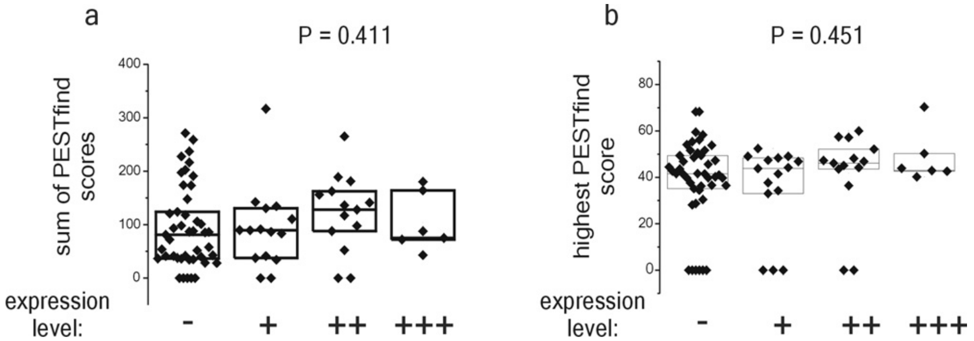

Fig. 5. Distribution of PEST motifs. PEST motifs as well characterized motifs enhancing protein degradation in vivo were determined using the program PESTfind. The output scores were converted into the positive range by adding 50 to allow easier processing. A score of zero was assigned to cDNAs giving no PESTfind hit. A score higher than 50 is by definition regarded as a putative PEST sequence. Plotted are the sum of all PESTfind scores along the whole proteins sequence (a) as well as only the highest score per protein (b). Neither of them is significantly associated with a level of expression.

3.4. General protein features

Looking at overall protein features, neither the protein length ($P = 0.575$; Fig. 4a), nor the aromaticity of the protein ($P = 0.607$; Fig. 4b), or the hydrophobicity of the protein ($P = 0.945$; Fig. 4c), correlate with the expression level. A significant association ($P = 0.029$, Fig. 4d) is, however, observed between a relatively high isoelectric point and non-detectable expression (median of 7.8 in contrast to medians from 6.0 to 6.3 of the other categories). A high p*I* seems to prevent even low protein yields.

3.5. Protein degradation signals

The PEST motif is a well characterized motif enhancing protein degradation in vivo (Rogers et al., 1986). This hydrophilic motif greater than or equal to 12 amino acids has been shown to confer instability to a range of proteins in a variety of species, including yeast. A considerable body of evidence supports the idea that PEST sequences target proteins for degradation by the 26S proteasome after ubiquitination (reviewed by Rechsteiner and Rogers, 1996). The algorithm PESTfind (Rechsteiner and Rogers, 1996) used here identifies PEST-like motifs and assigns a score depending on the degree of consensus of the particular motif with the requirements. After transforming the output scores into positive numbers (see Section 2), a score higher than 50 is by definition regarded as a putative PEST sequence, and a score higher than 55 sparks real interest (Rechsteiner and Rogers, 1996). If no PEST-like motif was found within a protein, a score of zero would be assigned to the sequence. As more than one motif can occur per protein, more than one score can be assigned per protein sequence. To our knowledge it is not known if a potential additive effect of more than one PEST motif per protein exists. Therefore, we took into account the sum of PESTfind scores for each sequence (Fig. 5a) as well as only the highest score – meaning the “best” motif – for each protein (Fig. 5b). An association between either the sum of PEST motifs within each protein sequence or the highest PESTfind score in each protein with the expression level is, however, not detectable ($P = 0.411$ and 0.451, respectively). There are putative PEST sequences in each of the four categories. Since the occurrence of these sequences does not associate with one of the expression levels, protein degradation in vivo as a possible obstacle for expression cannot generally be ascribed to PEST like motifs.

To take a more general approach towards protein degradation, lysine residues being the sites of a possible ubiquitination and subsequent degradation were taken into account. The number of lysines per protein (Fig. 6a), the number of lysine residues

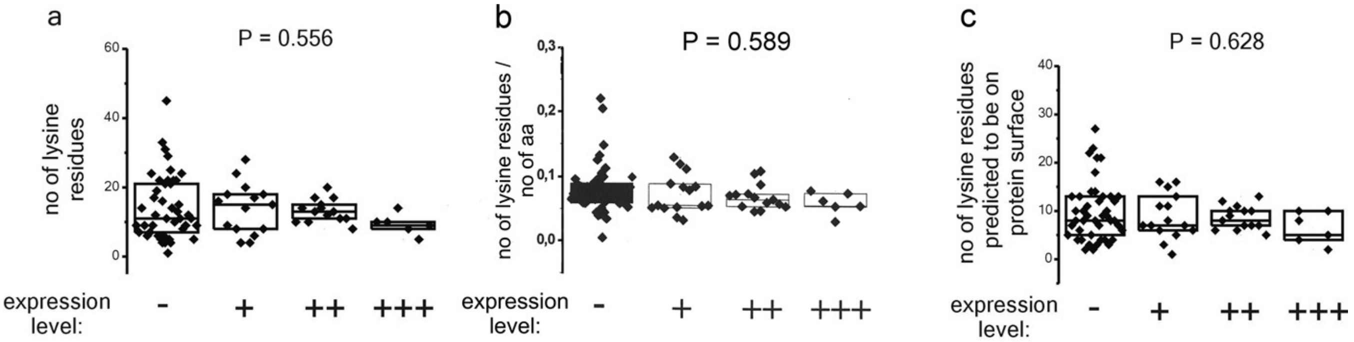

Fig. 6. Distribution of lysine residues. Lysine residues were taken into account as of their role as potential sites of ubiquitination and subsequent degradation of the protein. The surface probability of each lysine residue was taken into account to estimate the accessibility to the ubiquitination machinery. The surface probability was calculated using the Emini-Index by the program DNAStar. Lysine residues were assigned to have a high surface probability when the Emini-Index for the respective position was 1 or higher. Shown are the distributions of lysine residues per protein (a), the number of lysine residues per amount of amino acids of the respective protein (b) and lysine residues per protein with a surface probability of one or higher according to Emini (c). There is no correlation of either of those with the expression level.

per amount of amino acids of the respective protein (Fig. 6b) and the number of lysine residues predicted to be on the protein surface according to Emini et al., were analysed (see Section 2; Fig. 6c). Neither calculation revealed a statistically significant association with any of the expression categories ($P = 0.556$, 0.589 and 0.628, respectively). Therefore, occurrence of lysine residues – even under consideration of their surface probability to account for accessibility to the ubiquitination machinery – does not seem to be a major parameter influencing protein yield.

3.6. Similarities to yeast proteins

To see if a similarity to the endogenous proteins of the host has an influence on the expression level of a given protein, the occurrence of a homologue to our target proteins in yeast was checked. In all the categories of detectable protein expression, 49% of the target proteins have a homologue in yeast (50% for +++; 50% for ++; 47% for +; see Table 1). In contrast, in the category of no detectable protein expression, only 18% of the target proteins have a homologue in yeast. There is a significant ($P = 0.038$) association between the occurrence of a homologous protein in yeast and a positive expression result in *P. pastoris*.

4. Discussion

Our analysis reveals that parameters can indeed be identified that correlate with a detectable protein expression or with a certain expression level. These parameters seem to be influential on the expression level and consideration of these parameters will thus increase the probability of success of an expression experiment in *P. pastoris*.

While most of the sequence-based parameters analysed did not show a significant association with expression level, three independent features could be defined to have a statistically significant association with the expression level: (i) the rare occurrence of AT-rich regions in the cDNA is associated with a high expression level; (ii) a high isoelectric point is associated with no detectable protein expression; (iii) the occurrence of a homologous protein in yeast is associated with a general success of the expression experiment.

An important aspect to these results is the fact, that the cDNAs were preselected. Included in the study were proteins not larger than 500 amino acids, proteins with no structure deposited in the Protein Data Bank, proteins without predicted transmembrane domains, proteins without coiled-coil regions, and proteins with no compositional bias.

Of these limitations especially the size limitation might add bias to the results. So no conclusions regarding proteins larger than 500 amino acids can be made. Due to the focus of the study on intracellular expressed proteins, no putative transmembrane proteins are included. Since no structural features of the proteins or compositional bias within the protein sequence have been examined, these preselection criteria should have no impact on the results obtained here. The fact that all the tested proteins had no deposited structure at the time of selection should have no impact on the expression results in *P. pastoris* either.

The influence of the overall AT-content or of AT-rich regions on heterologous gene expression in *S. cerevisiae* (Milek et al., 2000; Romanos et al., 1991) as well as in *P. pastoris* (Gurkan and Ellar, 2003; Scorer et al., 1993) has been published in a number of studies. This AT-effect is due to a premature processing of the heterologous mRNA which might contain polyadenylation signals recognized in the yeast host but not in the donor organism (Romanos et al., 1991). Processing and polyadenylation signals in yeast are far less conserved than in mammals and AU-rich sequences easily fulfil yeast processing requirements, (Zhao et al., 1999) both with regard to the AU-rich signal motifs themselves and their order and spacing (Graber et al., 1999). Thus, processing of the messenger might occur at fortuitous sites not used in the original system. Due to the degenerate nature of the yeast processing signal, no useful assignments of these motifs to the sequences analysed in this study could be done.

Nevertheless, sequences showing a low amount of AT-clusters are present in all four categories, indicating this to be a necessary but not sufficient prerequisite for high protein yields.

Codon usage and its correlation with gene expression have extensively been studied in different organisms. In the yeast *S. cerevisiae*, codon usage varies considerably among different genes (Bennetzen and Hall, 1982; Sharp et al., 1986). Multivariant statistical analysis revealed a high degree of codon bias towards certain preferred codons and a high level of protein expression (Jansen et al., 2003; Sharp et al., 1986). The influence of codon usage on the expression level of individual heterologous proteins in yeast has been investigated. Low codon adaptation was made responsible for a low expression level in *S. cerevisiae* (Brocca et al., 1998; Li et al., 2003) and *P. pastoris* (Brocca et al., 1998; Sinclair and Choy, 2002).

On the other hand, a codon optimized gene expressed in *P. pastoris* that – as a consequence – had an increased GC-content showed a 10.6-fold increase in protein activity. A control construct with the same GC-content without an optimized codon adaptation led to a 7.5-fold increase in activity, indicating the GC-content to be the major contributor to increased protein activity (Sinclair and Choy, 2002). In several experiments analysing heterologous expression in *P. pastoris* the coding sequence of the target gene was changed to enhance the GC-content in order to eliminate fortuitous polyadenylation sites and to adapt codon usage at the same time (Gurkan and Ellar, 2003; Milek et al., 2000; Woo et al., 2002). Therefore, a discrimination of these parameters supposedly affecting protein expression is not possible.

It has to be mentioned that the CAI of the sequences analysed here is in general not very high (median of about 0.1 for all categories with no sequences having a significantly higher CAI, Fig. 2b). It is therefore not possible to evaluate the influence of a well adapted sequence from our data. However, the mean of the CAI of 6217 endogenous open reading frames in *S. cerevisiae* is 0.107 with a standard deviation of 0.017 (Coghlan and Wolfe, 2000). Therefore, the sequences analysed here are well within the range of naturally expressed yeast genes so that the codon adaptation should not hinder expression.

On the other hand, strong codon usage bias in *S. cerevisiae* is only observed for highly expressed genes (Jansen et al., 2003;

Sharp et al., 1986). Even for the relatively highly expressed proteins in this study, the expression level might still be too low to see an effect of low codon adaptation on the expression level.

Apart from the overall codon adaptation of a sequence, the occurrence of codons that are rare in *S. cerevisiae* has been analysed. Rare codons in *S. cerevisiae* have been found in a significantly higher frequency in unstable mRNAs than in stable ones (Herrick et al., 1990). The occurrence of unfavourable codons is not a major factor influencing expression yield either.

With respect to general protein features, an influence of a single one of these on protein yield has not been reported so far to our knowledge. Here, we report a significant association between a relatively high calculated isoelectric point of a protein and the lack of expression. The distribution of p*I* values in eukaryotic proteomes shows a trimodality (Schwartz et al., 2001) and this clustering is associated with the cellular localisation of the respective protein. Cytoplasmic proteins exhibit a distinct clustering around p*I* 5–6, integral membrane proteins around p*I* 8.5–9, and nuclear proteins are almost evenly distributed around the p*I* range. A similar clustering of cytoplasmic and membrane proteins associated with their localisation is observed in *Escherichia coli* (VanBogelen et al., 1999). There is no well-supported hypothesis to explain why cytosolic proteins should have p*I* values generally below 7 (Schwartz et al., 2001). Generally proteins in a pH value around their p*I* tend to aggregate. Since a non-native lysis and preparation procedure by boiling the cells in Laemmli-buffer (see Section 2) was used, total protein will be detected. Therefore, no difference has been made between aggregated and soluble protein.

Our observation might indicate a regulatory effect on protein expression or stability. It seems that the p*I* distribution associated with the cellular distribution did not only develop due to some evolutionary selection, but that the cell might somehow downregulate the cytosolic accumulation of proteins with a p*I* above a certain threshold.

When looking at the length of the proteins, it has to be considered that no proteins larger than 500 amino acids were included in the study. It is therefore not possible to draw any conclusions about the expression of large proteins.

A potential protein instability in vivo due to degradation via the proteasome was considered by looking for PEST like motifs and the occurrence of lysine residues as sites of ubiquitination. The latter search was refined by taking into account the surface probability of the respective lysine residue. Neither of these analyses showed any association with the protein expression level. Even if protein degradation was a major parameter for protein yield, an impact of these two parameters would not necessarily be detectable. PEST like motifs and ubiquitination do not account for all observed protein instabilities in vivo (Hoyt et al., 2003).

An association between the success of heterologous expression in a given host and the occurrence of a homologous protein in that host has not been described before. It is known that there are mechanisms in yeast that distinguish correctly folded and unfolded proteins and subsequently lead to the degradation of the latter (Welihinda et al., 1999). These mechanisms are also known to have an impact on the heterologous expression of proteins targeted to the secretory pathway (e.g. Hohenblum et al., 2004; Kauffman et al., 2002). The observed higher expression level of proteins having homologues in yeast could be due either to superior folding of these “yeast-like” proteins in yeast as the yeast folding apparatus is well adapted to the yeast protein set. Another explanation could be that homologous proteins are associated and thus stabilized in protein complexes. The major portion of the proteins in a yeast cell is part of larger complexes (Gavin et al., 2002). Integration of heterologous proteins due to their homology to the endogenous members of a complex might lead to a stabilization of the protein in the cell and therefore a higher expression level.

Interestingly, a much more general analysis of the data present in TargetDB, a database containing target proteins and their respective progress within most of the structural genomics networks, identified parameters associated with a positive protein expression (Goh et al., 2004). This analysis does not consider parameters like the origin of the protein (pro- or eukaryotic), expression host, expression construct, experimental procedure, etc., but since by far most of these projects use *E. coli* as expression host, the data are mostly relevant for this host. But regardless of this major difference, as the one parameter most strongly correlated to expression, conservation of the protein among organisms was identified. This observation supports our observation regarding the expression success of yeast-homologous human proteins in *P. pastoris*. Other parameters associated with expression success identified by Goh et al. (2004) are protein length, protein hydrophobicity, percentage of small negatively charged residues and – ranking fifth in importance – the isoelectric point. This observation indicates that the p*I* might indeed be a general parameter influencing protein expression. This is consistent with the above mentioned clustering of p*I*s with their intracellular localisation.

Probably there will be more parameters influencing the success of an expression experiment than were analyzed within this study. Further parameters to look at could be secondary structure at the translation initiation region, GC-cluster or G-clusters. Clusters of rare codons could also influence the expression. Further work on a larger sample of tested cDNAs can reveal more influential parameters.

We show that a sequence-based approach alone can identify parameters associated to heterologous expression results. The results presented here might contribute to eventually providing information towards the choice of *P. pastoris* as a suitable host system and the knowledge-based optimisation of gene sequences to increase the protein yield. The results presented here in combination with more similar information should prove valuable to save time and costs for expression trials and design high-throughput approaches.

Acknowledgements

We thank Ulrich Harttig for help with tblastn and PERL programming. Natalia Bolotina amplified some of the cDNAs.

We thank Roslin Bensmann for help with the English. Our colleagues in the protein structure factory project, especially Bianka Prinz and Caterina Holz provided helpful discussion. This work was supported by the Bundesministerium für Bildung und Forschung through the Leitprojekt "Proteinstrukturfabrik" to CL.

References

Bennetzen, J.L., Hall, B.D., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026–3031.

Boettner, M., Lang, C., 2004. High-throughput expression in microplate format in *Pichia pastoris*. Methods Mol. Biol. 267, 277–286.

Boettner, M., Prinz, B., Holz, C., Stahl, U., Lang, C., 2002. High-throughput screening for expression of heterologous proteins in the yeast *Pichia pastoris*. J. Biotechnol. 99, 51–62.

Brocca, S., Schmidt-Dannert, C., Lotti, M., Alberghina, L., Schmid, R.D., 1998. Design, total synthesis, and functional overexpression of the Candida rugosa lip1 gene coding for a major industrial lipase. Protein Sci. 7, 1415–1422.

Brooksbank, C., Camon, E., Harris, M.A., Magrane, M., Martin, M.J., Mulder, N., O'Donovan, C., Parkinson, H., Tuli, M.A., Apweiler, R., Birney, E., Brazma, A., Henrick, K., Lopez, R., Stoesser, G., Stoehr, P., Cameron, G., 2003. The European Bioinformatics Institute's data resources. Nucleic Acids Res. 31, 43–50.

Coghlan, A., Wolfe, K.H., 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast 16, 1131–1145.

Comeron, J.M., Aguade, M., 1998. An evaluation of measures of synonymous codon usage bias. J. Mol. Evol. 47, 268–274.

Emini, E.A., Hughes, J.V., Perlow, D.S., Boger, J., 1985. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J. Virol. 55, 836–839.

Garber, K., 2001. Biotech industry faces new bottleneck. Nat. Biotechnol. 19, 184–185.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147.

Goh, C.S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G.T., Zhao, H., Gerstein, M., 2004. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. J. Mol. Biol. 336, 115–130.

Graber, J.H., Cantor, C.R., Mohr, S.C., Smith, T.F., 1999. Genomic detection of new yeast pre-mRNA 3′-end-processing signals. Nucleic Acids Res. 27, 888–894.

Gurkan, C., Ellar, D.J., 2003. Expression of the Bacillus thuringiensis Cyt2Aa1 toxin in *Pichia pastoris* using a synthetic gene construct. Biotechnol. Appl. Biochem. 38, 25–33.

Heinemann, U., Frevert, J., Hofmann, K., Illing, G., Maurer, C., Oschkinat, H., Saenger, W., 2000. An integrated approach to structural genomics. Prog. Biophys. Mol. Biol. 73, 347–362.

Herrick, D., Parker, R., Jacobson, A., 1990. Identification and comparison of stable and unstable mRNAs in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 10, 2269–2284.

Hohenblum, J., Gasser, B., Maurer, M., Borth, N., Mattanovich, D., 2004. Effects of gene dosage, promotors, and substrates on unfolded protein stress of recombinant *Pichia pastoris*. Biotechnol. Bioeng. 85, 367–375.

Holz, C., Prinz, B., Bolotina, N., Sievert, V., Büssow, K., Simon, B., Stahl, U., Lang, C., 2003. Establishing the yeast *S. cerevisiae* as a system for expression of human proteins on a proteome-scale. J. Struct. Funct. Genomics 4, 97–108.

Hoyt, M.A., Zhang, M., Coffino, P., 2003. Ubiquitin-independent mechanisms of mouse ornithine decarboxylase degradation are conserved between mammalian and fungal cells. J. Biol. Chem. 278, 12135–12143.

Jansen, R., Bussemaker, H.J., Gerstein, M., 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res. 31, 2242–2251.

Kauffman, K.J., Pridgen, E.M., Doyle 3rd, F.J., Dhurjati, P.S., Robinson, A.S., 2002. Decreased protein expression and intermittent recoveries in BiP levels result from cellular stress during heterologous protein expression in *Saccharomyces cerevisiae*. Biotechnol. Prog. 18, 942–950.

Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assoc. 47, 583–621.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Lennon, G., Auffray, C., Polymeropoulos, M., Soares, M.B., 1996. The I. M. A. G. E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics 33, 151–152.

Li, H., Ma, Y., Su, T., Che, Y., Dai, C., Sun, M., 2003. Expression, purification, and characterization of recombinant human neurturin secreted from the yeast *Pichia pastoris*. Protein Exp. Purif. 30, 11–17.

Milek, R.L., Stunnenberg, H.G., Konings, R.N., 2000. Assembly and expression of a synthetic gene encoding the antigen Pfs48/45 of the human malaria parasite Plasmodium falciparum in yeast. Vaccine 18, 1402–1411.

Rechsteiner, M., Rogers, S.W., 1996. PEST sequences and regulation by proteolysis. Trends Biochem. Sci. 21, 267–271.

Ribeiro, J.M., Sillero, A., 1991. A program to calculate the isoelectric point of macromolecules. Comput. Biol. Med. 21, 131–141.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.

Rogers, S., Wells, R., Rechsteiner, M., 1986. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science 234, 364–368.

Romanos, M.A., Makoff, A.J., Fairweather, N.F., Beesley, K.M., Slater, D.E., Rayment, F.B., Payne, M.M., Clare, J.J., 1991. Expression of tetanus toxin fragment C in yeast: gene synthesis is required to eliminate fortuitous polyadenylation sites in AT-rich DNA. Nucleic Acids Res. 19, 1461–1467.

Schwartz, R., Ting, C.S., King, J., 2001. Whole proteome p*I* values correlate with subcellular localizations of proteins for organisms within the three domains of life. Genome Res. 11, 703–709.

Scorer, C.A., Buckholz, R.G., Clare, J.J., Romanos, M.A., 1993. The intracellular production and secretion of HIV-1 envelope protein in the methylotrophic yeast *Pichia pastoris*. Gene 136, 111–119.

Sharp, P.M., Cowe, E., 1991. Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast 7, 657–678.

Sharp, P.M., Li, W.H., 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Sharp, P.M., Tuohy, T.M., Mosurski, K.R., 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14, 5125–5143.

Sinclair, G., Choy, F.Y., 2002. Synonymous codon usage bias and the expression of human glucocerebrosidase in the methylotrophic yeast, *Pichia pastoris*. Protein Exp. Purif. 26, 96–105.

Stevens, R.C., 2000. Design of high-throughput methods of protein production for structural biology. Struct. Fold. Des. 8, R177–R185.

VanBogelen, R.A., Schiller, E.E., Thomas, J.D., Neidhardt, F.C., 1999. Diagnosis of cellular states of microbial organisms using proteomics. Electrophoresis 20, 2149–2159.

Weissman, A.M., 2001. Themes and variations on ubiquitylation. Nat. Rev. Mol. Cell. Biol. 2, 169–178.

Welihinda, A.A., Tirasophon, W., Kaufman, R.J., 1999. The cellular response to protein misfolding in the endoplasmic reticulum. Gene Exp. 7, 293–300.

Woo, J.H., Liu, Y.Y., Mathias, A., Stavrou, S., Wang, Z., Thompson, J., Neville Jr., D.M., 2002. Gene optimization is necessary to express a bivalent antihuman anti-T cell immunotoxin in *Pichia pastoris*. Protein Exp. Purif. 25, 270–282.

Wright, F., 1990. The ‘effective number of codons’ used in a gene. Gene 87, 23–29.

Yokoyama, S., 2003. Protein expression systems for structural genomics and proteomics. Curr. Opin. Chem. Biol. 7, 39–43.

Zhang, S.P., Zubay, G., Goldman, E., 1991. Low-usage codons in *Escherichia coli*, yeast, fruit fly and primates. Gene 105, 61–72.

Zhao, J., Hyman, L., Moore, C., 1999. Formation of mRNA 3’ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol. Mol. Biol. Rev. 63, 405–445.