UDC 57.081.23
ISSN 1330-9862

review

(FTB-1113)

Vector Design Related to Bio-Processing

*Reingard Grabherr and Karl Bayer**

Institute of Applied Microbiology, University of Agricultural Sciences, Muthgasse 18, A-1190 Vienna, Austria

Received: October 30, 2001
Accepted: November 8, 2001

Summary

Increased performance of informatics, molecular biological tools and engineering have led to new concepts and scopes in recombinant gene expression technology. To take full advantage of an integrated production process, molecular biology, cell biology and engineering aspects have to be considered in order to gain an efficient and stable process. The contribution of the individual key factors, such as transcriptional and translational activity, codon usage and gene dosage is being assessed and evaluated, with respect to bio-processing. The rapid development of high through-put and complex analytical methods deliver genome and proteome wide data that can be used to construct regulatory networks. This greatly contributes to the understanding of cellular processes and helps to reveal metabolic bottlenecks during production of the individual recombinant protein. The impact of design of the genetic construct is being high-lighted, as it is the initial step in bio-processing, and must contain all relevant features for optimal and mutual exploitation of the host.

Key words: heterologous gene expression, transcriptional regulation, translational regulation

Introduction

Whatever the goal for molecular biology research and development is, be it molecular tools, recombinant gene production, it is the genetic construct embedded in the constraints of host metabolism which is the initial determinant for bio-processing. In order to facilitate and enable prediction of process design, the features of the expression vector are of crucial importance. In addition to economic aspects, vector design is strongly determined by stringent regulatory issues with respect to processing. Therefore, short-term set-up of an optimized gene construct is a key issue in bio-process development. Operation procedures have to be laid down very early and further steps, such as downstream processing and clinical trials, depend on the early availability of the recombinant protein in sufficient amounts and reproducible quality.

The key demands of an optimized production process are:

- rapid cloning and screening methods to match the speed of scientific progress in the field of life sciences
- optimal control and exploitation of the host cell metabolic bio-molecule synthesis machinery
- adaptation of process units for efficient production
- fulfillment of safety and regulatory requirements

A. The need of vector design

Controlled regulation of heterologous gene expression is of major importance for a wide variety of basic and applied biological research and production areas,

* Corresponding author; Phone: ++ 43 1 36006 6242; Fax: ++ 43 1 369 7615; E-mail: Bayer@mail.boku.ac.at

including manufacture of biopharmaceuticals, functional genomics, tissue engineering and gene therapy. Recombinant protein expression technology is based on the interplay of gene regulation, the metabolic properties of a chosen host, optimized culture conditions, production process control, purification strategies, analytics and quality control. The final product, as it is desired, is the result of a complex network of interacting factors and parameters. Each manipulation of a single factor influences the entire complex of the cell factory. Design of the expression vector, containing the genetic construct, has to be revisited, even more so, since fermentation strategies have improved and their concept has changed during the past years. The availability of computing has greatly enhanced bio-process engineering capabilities. For bacterial fermentation, fed-batch strategies are now state of the art, whereby nutrients are being added continuously, according to biomass increase (BTM), growth rate (μ) and/or production rate (qP). Further, the demand of novel products, *e.g.* high quality plasmid DNA instead of protein or even entire reaction cascades, requires targeted adaptation of the whole process design. Modern molecular biology has generated novel, powerful tools, such as micro arrays for transcriptome analysis, differential display techniques and fast and efficient PCR and sequence strategies for high-throughput screening. Fluorescence activated cell sorters and two-dimensional electrophoresis have further extended the analytical repertoire in molecular biology. Hence, a new era of bio-processing technology has begun, and new requirements can be fulfilled, exploiting new insights, novel tools, and rational design concepts.

B. The basics of vector design

Dynamic flexibility defines any cellular organism and is a prerequisite in evolution. Highly sophisticated regulatory metabolic circuits must be able to react to environmental changes. In contrast to »predictable« challenges and variations, when a genome is confronted by a burden for which it is unprepared it may reorganize itself, *e.g.* induce shift in metabolic enzyme activity, alter the balance between stability and repair, increase genetic variation and change the spectrum of mutations. For recombinant gene expression we have to broaden our understanding for the entity of the system and study single steps of regulation intensively in order to manipulate and redirect the cell factory in the desired way. Each change of the host's genetic structure, influences the whole organism in various aspects, and causes events, which again affect the genetic substance. It has been demonstrated that starvation-induced derepression results in enhanced mutation rates by stimulating rates of transcription in targeted operons, thereby increasing the concentration of single-stranded DNA, which is more vulnerable to mutations than double-stranded DNA. (*1*,*2*).

The development of cloning tools and techniques starting in the mid-70s, the main focus was to optimize the genetic construct in respect to maximal transcription and translation efficiency, without considering the host's capacity and its triggered response mechanisms. Though genetic engineering promises better and more plentiful products, genetically engineered organisms may encounter unpredictable obstacles, such as hampered process control. Using the potential and synergies of novel tools such as genomics, proteomics, and metabolomics, the very complex nature of intracellular networks becomes evident. If we succeed in understanding, exploiting and connecting these highly specific data, a major step forward in bio-processing and its applications would be achieved (*3*).

Bioinformatics, metabolic engineering and inducible expression systems have provided the tools for rational design of production systems. Based on these tools an integrated system approach must be applied. The first module is the genetic construct which has to be designed, in accordance with host properties, process control and product requirements.

Regulating Heterologous Gene Expression

A. Transcriptional regulation

We now understand transcriptional initiation control to be largely the result of transcription factor complexes interacting with RNA polymerase to inhibit or stimulate transcription from a given promoter. Proper timing and levels of transcription are controlled by interaction of one or more transcription factors with external or internal signals, sometimes directly or indirectly by means of a signal transduction cascade. Basically, transcription can be regulated endogenously by the use of an inducable promoter, *e.g*. Lac- and Tac-promoter in *E. coli* (*4*,*5*) methanol inducible promoters in yeast (for review see *6*) or heat shock promoters in mammalian cells (*7*, *8*, for review *9*). Besides the rate of mRNA production, its stability is of major relevance for protein production and the metabolic exploitation of the host (*10*,*11*).

From an engineering point of view, bacterial plasmid encoded recombinant protein production offers the greatest advantages, due to high yield, controllability and efficiency. Depending on the objectives to be achieved either broad host or narrow host range plasmid species with different replication systems, which lead to low, medium and high copy number plasmids, are used. In recent years a trend towards the use of strong promoter systems to increase product yield can be recognised. Although remarkable efforts could be achieved (for review see *12*), these solutions mainly based on genetics showed also severe drawbacks. The too high transcription rate of the target mRNA triggers a metabolic overload of the host cell and curtails the synthesis of mRNA translated into cellular proteins (*13*). Therefore, the promoter system must be designed to enable control of recombinant gene expression in order to establish a tolerable equilibrium between recombinant and host cell protein formation. Hence, ways to attain the appropriate recombinant gene expression rate are modulation of transcription rate and/or variation of the gene dosage. Control of transcription rate can be rather easily accessed by limiting amounts of inducer, whereas control of gene dosage is far more difficult to regulate continuously (*14*).

Remarkably, the impact of the induction process has not been recognised as a priority aspect in research in the past. As a standard operation the inducer is supplied by a pulsed feed, thereby triggering full induction,

which leads in most cases to over-expression and does not allow modulation of transcription rate. An alternative concept uses inducer titration by feeding non-saturating amounts of inducer in a constant ratio to the increase of biomass (*15*). This implies that growth must proceed under predefined steady state conditions, which are accomplished in fed-batch fermentation applying an exponential feed algorithm. To gain the full potential of inducer titration and transcription rate control, the interaction of inducer with host cell metabolism has to be taken into account. In assumption that cellular growth follows Monod-kinetics, the efficiency of inducer transport into the cell (*16*), is strongly influenced by the overall availability of substrate, thus transcription rate is a function of the environmental conditions. Chemostat culture experiments, using the lac system, showed that the amount of inducer (IPTG) had to be increased at increasing growth rates (*17*). This phenomenon is even more pronounced with inducers, which are delivered into the cell by specific transport proteins, which is not the case with the gratuitous inducer IPTG. These transport mechanisms in combination with inducer limited transcription rate may even lead to the evolution of an over-producing and a non-producing partition of the population and thereby alter the behaviour of the total population (*18*). As a whole the design of a sophisticated induction strategy with a comprehensive view on host cell properties is evident.

B. Gene copy number

B.1. Need of regulation

In addition to the rate of transcription and stability of mRNA, the target gene dosage implies an important impact on the specific mRNA level, and hence on the host's metabolic regulatory mechanisms. While gene copy number in mammalian cells can be raised by including a drug resistance gene (*e.g.* dihydrofolate reductase gene) and adding an increasing dose of this drug (*e.g.* methotrexate), in yeast there is usually one copy of the desired gene, which has been introduced by homologous recombination into a specified locus on the genome. In bacteria, high copy number plasmids yield between 40 and 300 copies of the target gene per cell. This gives the advantage of a substantial amount of foreign DNA, present in the host, which drastically reduces the probability of plasmid loss during growth. High replication rates, however, can increase to a level, where the cellular control mechanisms fail, such as stringent control, and the metabolic capacity of the host is overstrained. Therefore, bacterial systems, unlike others, have to be regulated concerning the heterologous gene dosage (*19*). Increase in plasmid replication causes additional stress within the host because not only does replication increase, but consequently the transcriptional and the translational machinery are extremely challenged and soon run out of metabolic building blocks and energy. It could be shown in fed-batch fermentation with controlled feed of inducer, that the expression of recombinant protein could be kept below a critical value, as described above. However, to achieve full controllability of trancriptional regulation, stabilisation of the plasmid replication within a desired range must be provided by modification of the relevant control elements. There are three general types of bacterial plasmid copy number control systems. Either directly repeated sequences (iterons) form a complex with cognate replication initiator proteins, or antisense RNA binds to proteins or to a complementary RNA primer (*20*). The most widely used plasmids are derivatives of the plasmid pBR322. Its origin of replication, *Col*E1, is controlled by an anti-sense RNA, which inhibits the maturation of the primer, essential for replication, by binding.

To attain bacterial process optimisation, several attempts to change the plasmid copy number have been reported. Replication rates could be modified by incorporating a point mutation that affects initiation of replication whereby a pBR322-derived plasmid could be converted into a high copy number plasmid (*21*,*22*). Further results suggest that mutating the promoters of RNA I and/or RNA II is a possible option to influence plasmid copy number (*23*).

A novel approach to adjust replication rates is to directly target the binding mechanism of the two major regulators of plasmid replication and inhibition, RNA I and RNA II. The major objectives are to either maintain and/or to restore the function of the replication control mechanism, or to increase replication by exploiting the relevant regulatory elements. Based on the finding that uncharged tRNAs disturb the replication control mechanism (*24*), deliberately decreasing the homology of tRNAs to the corresponding loops of RNA I and RNA II can be expected to diminish the binding of tRNAs, and provide appropriate hybridization of RNA I and RNA II. Thus the replication control system remains functional in spite of high tRNA levels derived from high expression rates (*14*).

C. Translational regulation

Translation is a key step in protein synthesis, determining the order of amino acids in a protein. Thus, translation must be relatively error free in order to allow the accurate flow of genetic information (for review see *25*). Regulatory mechanisms have developed in order to adjust and fine-tune protein synthesis during the steps involved in translation. In bacteria, mRNA translation is modulated by translation attenuation, whereby a specific ribosome binding site is being sequestered by mRNA secondary structure (*26*,*27*). For optimal translation efficiency, sequence specific features can be included (*28*), according to the particular organism. In addition to the Shine-Dalgarno sequence and the start codon, other sequences in the mRNA have been found to be important for efficient translation (*12*). Dong *et al.* found that over-expression of gratuitous proteins from high copy number plasmids leads to destruction of ribosomes (*29*). Although the mechanisms involved are not clear, the impact of these findings must be considered in bio-process design, and should be further investigated using genome-wide analysis.

D. Codon usage

Examining the codon usage of various organisms revealed that not all 61 mRNA codons are used equally. Codon usage differs among genomes (*30*), among different

genes within the same genome (*31,32*), and even among different segments of the same gene (*33,34*). Usually, the frequency of the codon usage reflects the abundance of their cognate tRNAs. Therefore, when the codon usage of a target gene differs significantly from average codon usage of the expression host, problems are often encountered, such as decreased mRNA stability, premature termination of transcription and/or translation, frameshifts, deletions and misincorporations. Based on the data available, computer programs have been developed (*35*) in order to modify the gene of interest accordingly. The use of yeast-preferred codons resulted in a more than 5-fold increase in the rate of synthesis and at least a 50-fold increase in the steady state level of protein (*36*). Improvement of expression has further been shown when human genes are to be expressed in *E. coli* (*e.g. 37*).

Contradictory conclusions have been drawn concerning the relationship between codon usage and secondary structure. In 1996, Thanaraj and Argos stated that different protein secondary structural types are differentially coded on mRNA in *E. coli*. (*38*). While Xie *et al*. (*39*) stated that there exists a correlation in mammals but not in prokaryotes, Gupta *et al*. (*40*) found that instead of the codon usage, it is the occurrence of bases at the second codon position that differs in secondary structural units in prokaryotes and eukaryotes. Transferring a gene from one organism to another influences the new host, as codon usage and amino acid sequence differ between organisms. Codon optimization must be performed in consideration of the genes' descendence, the host's codon usage and availability of specific tRNAs, the biased codon usage of host and donor, and eventual protein structure domains. Extended data bases and sophisticated algorithms are the basis of successful sequence adaptation and will be needed for the purpose of expression optimization. Although we are yet far from full understanding, the code which is embedded in the RNA, respectively the codon sequence, can be viewed as the information that sets the rules for complex and interacting regulatory steps in the cells life cycle.

Conclusions and outlook

There is a long way to go until we fully understand all single steps of cellular mechanisms and their interactive roles for recombinant gene expression technology. We have to investigate and view interactions and circuits as parts of entities, rather than looking at them as single components and steps, since regulatory mechanisms are always part of a hierarchically ordered and interactive structure. All the information about the impact on the host, metabolic interactions and physical properties of the product, are manifested in the DNA sequence of the genetic construct. Every new piece of information about the organism, such as its stress response, codon usage, transcriptional behavior, ribosomal binding preferences and nutrient demands must flow back into the genetic design of the expression vector. This design is of high value only when an optimized host-vector system is able to perform at full potential and consequently serves to improve manipulation and regulation of the entire expression system.

The individual components required for efficient expression have been investigated in detail and are well described by molecular biologists. However, their assembly to achieve optimal functionality lacks some systematic approach, although empirical rules have emerged.

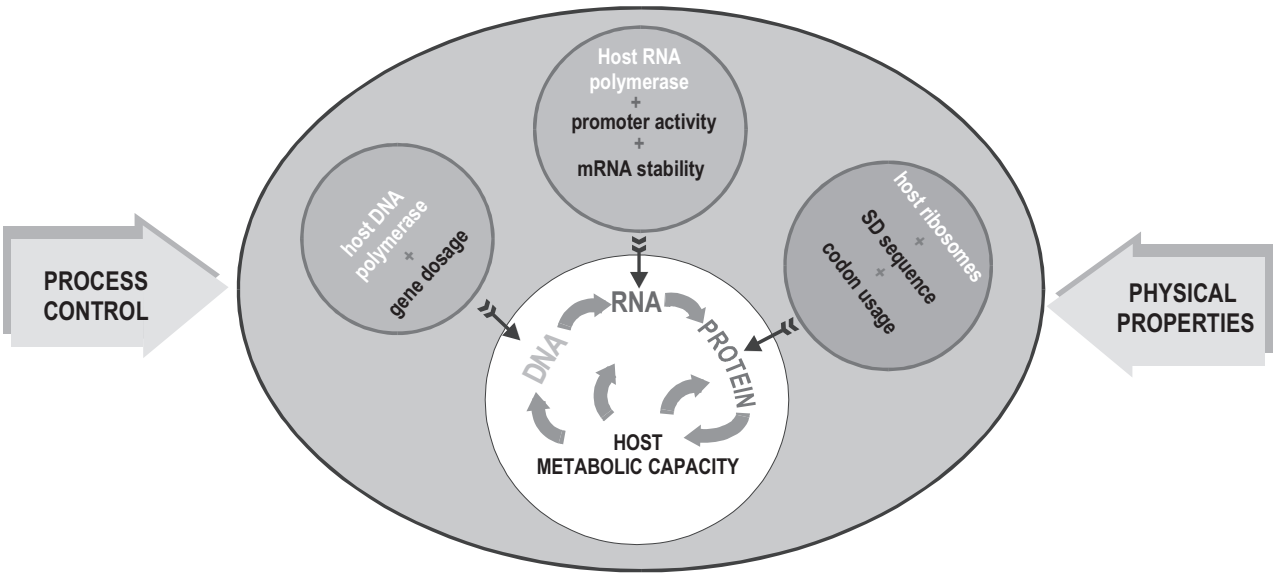

Fig. 1. Scheme describes the complex, influential factors on host capacity and their interaction inside and outside a bio-processor. The host DNA polymerase and other DNA replicating enzymes (host DNA polymerase), in combination with the gene dosage results in the production of DNA. The host transcriptional machinery (host RNA polymerase) together with the specific promoter activity and the stability of the derived transcript, is responsible for the production of specific mRNA. The host ribosomal machinery (host ribosomes), the Shine-Dalgarno-sequence and other translational signals (SD sequence) are responsible for the production of specific protein. All three products (DNA, RNA, protein) interact with the host's metabolic capacity. Process control and physical properties of the fermentor are further factors influencing the production process from outside.

While protocols for control of transcriptional and translational regulation have been developed, the effects and interactions with the cellular processes are yet less understood. The primary prerequisite for investigating the behavior of expression vectors and its interaction with the host is the reproducible, stable and defined operation of bio-processes. Only in this case it is permissible to acquire data using the extended analytical repertoire such as genomic, proteomic and metabolomic tools. Derived from these results, the functions, cooperative networking, interference and limits of the involved metabolic pathways can be visualized and redirected for chosen processes. Nowadays, process development and design benefits enormously from the increased performance of informatics, biology and engineering and its interdisciplinary integration. All thereof derived information must be contained in the genetic vector construct.

References

1. B. E. Wright, *FEBS Lett. 402* (1997) 4.
2. B. E. Wright, A. Longacre, J. M. Reimers, *Proc. Natl. Acad. Sci. USA, 96* (1999) 5089.
3. M. W. Covert, C. H. Schilling, I. Famili, J. S. Edwards, I. I. Goryanin, E. Selkov, B. O. Plasson, *Trends Biochem. Sci. 26* (2001) 179.
4. D. T. Denhardt, J. A. Colasanti, *Biotechnology, 10* (1988) 179.
5. R. S. Donovan, C. W. Robinson, B. R. Glick, *J. Ind. Microbiol. 16* (1996) 145.
6. J. L. Cereghino, J. M. Cregg, *FEMS Microbiol. Rev. 24* (2000) 45.
7. R. Holmgren, K. Livak, R. Morimoto, R. Freund, M. Meselson, *Cell, 18* (1979) 1359.
8. F. M. Wurm, K. A. Gwinn, R. E. Kinston, *Natl. Acad. Sci. USA, 83* (1986) 5414.
9. M. Fussenegger, *Biotechnol. Progr. 17* (2001) 1.
10. T. A. Carrier, J. D. Keasling, *Biotechnol. Progr. 13* (1997) 699.
11. T. A. Carrier, J. D. Keasling, *Biotechnol. Progr. 15* (1999) 58.
12. S. C. Makkrides, *Microbiol. Rev. 60* (1996) 512.
13. S.-T. Liang, Y. C. Xu, P. Dennis, H. J. Bremer, *J. Bacteriol. 182* (2000) 3037.
14. R. Grabherr, E. Nilsson, G. Striedner, K. Bayer, *Biotechnol. Bioeng.* (2001) in press.
15. G. Striedner, M. Cserjan-Puschmann, R. Grabherr, F. Clementschitsch, E. Nilsson, K. Bayer: *Recombinant Protein Production with Prokaryotic and Eukaryotic Cells. A Comparative View on Host Physiology*, O. W. Merten, D. Mattanovich, C. Lang, G. Larsson, P. Neubauer, D. Posso, P. Postma, J. Teixeira deMattos, J. Cole (Eds.), Kluwer academic publishers (2001).
16. A. Death, T. J. Ferenci, *J. Bacteriol. 176* (1994) 5101.
17. G. Striedner, H. Reischer, F. Pötschacher, K. Bayer, *Appl. Microbiol. Biotechnol.* (submitted for publication).
18. J. D. Keasling, *TIBTECH, 17* (1999) 452.
19. R. Grabherr, K. Bayer, *Trends Biotechnol.* (2001) in press.
20. G. DelSolar, M. Espinosa, *Mol. Microbiol. 37* (2000) 492.
21. R. Lahijani, G. Hulley, G. Soriano, N. A. Horn, M. Marquet, *Hum. Gene Ther. 7* (1996) 1971.
22. A. G. Bert, J. Burrows, C. S. Osborne, P. N. Cockerill, *Plasmid, 44* (2000) 173.
23. Y.-L. Yang, B. Polisky, *Plasmid, 41* (1999) 55.
24. B. Wróbel, G. Wegrzyn, *Plasmid, 39* (1998) 48.
25. M. Ibba, D. Söll, *Science, 186* (1999) 1893.
26. P. S. Lovett, *Gene, 179* (1996) 157.
27. C. Yanofsky, *J. Bacteriol. 182* (2000) 1.
28. C. M. Stenstrom, H. Jin, L. L. Major, W. P. Tate, L. A. Isaksson, *Gene, 263* (2001) 273.
29. H. Dong, L. Nilsson, C. G. Kurland, *J. Bacteriol. 177* (1995) 1497.
30. R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pave, *Nucleic Acids Res. 8* (1980) 49.
31. M. Gouy, C. Gautier, *Nucleic Acids Res. 19* (1982) 7055.
32. H. Grosjean, W. Fiers, *Gene, 18* (1982) 1999.
33. H. Akashi, *Genetics, 136* (1994) 972.
34. X. Xia, *Genetics, 149* (1998) 37.
35. E. Wolf, P. S. Kim, *Protein Sci. 8* (1999) 680.
36. L. Kotula, P. J. Curtis, *Biotechnology, 9* (1991) 1386.
37. R. S. Hale, G. Thompson, *Protein. Expr. Purif. 12* (1998) 185.
38. T. A. Tharanaraj, P. Argos, *Protein Sci. 5* (1996) 1973.
39. T. Xie, D. Dind, X. Tao, D. Dafu, *FEBS Lett. 434* (1998) 93.
40. S. K. Gupta, S. Majumdar, T. K. Bhattachary, T. C. Gosh, *Biochem. Biophys. Res. Com. 269* (2000) 692.

Konstrukcija vektora za bioprocesiranje

Sažetak

Dostignuća u informatici, molekularno-biološkim tehnikama i inženjerstvu dovela su do novih shvaćanja i ciljeva u tehnologiji rekombinantne genske ekspresije. Kako bi se postigao puni probitak u integriranom proizvodnom procesu, potrebno je uvažiti sve aspekte molekularne biologije, biologije stanice i inženjerstva da bi se postigao djelotvoran i stabilan proces. Utvrđen je i procijenjen doprinos pojedinih ključnih činitelja, kao što su transkripcijska i translacijska aktivnost, uporaba kodona i količine gena na tijek bioprocesiranja. Brzi razvoj vrlo učinkovitih i kompleksnih analitičkih postupaka daju opširne podatke o genomu i proteomu koji se mogu koristiti za konstrukciju regulacijske mreže. To bitno pridonosi razumijevanju staničnih procesa i pomaže u otkrivanju metabolički uskih grla tijekom proizvodnje pojedinog rekombinantnog proteina. Osobito je istaknut utjecaj nacrta genetske konstrukcije, jer je on početni korak u bioprocesiranju koji mora sadržavati sve relevantne značajke za optimalnu i uzajamnu eksploataciju domaćina.