Construction of an *EPO* (*Human-Erythropoietin*) Synthetic Gene Through a *Recursive*-PCR Method

Asrul Muhamad Fuad[1*], Tutus Gusdinar[2], Debbie Sofie Retnoningrum[2], and Dessy Natalia[3]

[1]Research Center for Biotechnology – Indonesian Institute of Sciences (LIPI) Cibinong Science Center, Cibinong
[2]School of Pharmacy and [3]Dept. of Chemistry - Faculty of Mathematics and Life Sciences, Institute Technology Bandung, Bandung
**asrul.fuad@yahoo.co.id*

ABSTRACT

Human erythropoietin (hEPO) is an important glycoprotein in human that is coded by a single gene named *EPO* (*eryhtropoietin*). EPO is a glycoprotein hormone that promotes erythropoiesis, which is the formation process of mature red blood cells (erythrocytes) in human bodies. It is widely used for treatment of anemia in patients with chronic renal failure. Therefore EPO has been classified as hematopoietic cytokine. Recombinant hEPO (rhEPO) has been commercially available, such as Epogen. It is produced in mammalian cells, such as CHO (Chinese hamster ovary) cells for the reason of its complex structure as a glyco-protein. In an effort to use and optimize heterologous *EPO* gene expression in an alternative eukaryotic host cells such as yeast, an *EPO*-synthetic gene (*EPOsyn*) was constructed. The synthetic gene had been designed to contain optimal *Pichia pastoris* codon usage. It had been constructed by a *recursive*-PCR method in two-step PCR reactions. The gene was assembled from 8 single strands synthetic oligonucleotides having an average length of 90 nt with 20 to 30 overlap region between two adjacent oligos. The synthetic gene has less GC content (45.31%) compared its native (human) gene (59.08%). The synthetic gene has been cloned in pCR2.1 cloning plasmid and sequenced. From 8 independent clones, it was revealed that the error rate was 1.59%, in which 1.42% was due to deletions and 0.17% due to substitutions. Design of the gene sequences, construction method and DNA sequence analysis of the gene will be discussed in this paper.

Keywords: Human erythropoietin (hEPO), erythropoiesis, *EPO*-synthetic gene, *recursive*-PCR, *Pichia pastoris*, hematopoietic cytokine.

Introduction

Human-erythropoietin (EPO) is one of hematopoietic growth factors that promotes erythropoiesis, which is the proliferation and differentiation of erythroid progenitor cells into erythrocytes (red blood cels) (Koury and Bondurant, 1992). EPO is a cytokines that is produced mainly in kidney and liver. EPO is produced as a response to low oxygen concentration in blood known as hypoxia (Yin and Blanchard, 2000). Other than in kidney and liver, EPO is also produced by neuroblastoma cells induced by hypoxia (Stolze *et al*, 2002) and in brain after oxidative or nitrosative stress (OMIM 133170). Sakanaka *et al* (1998) reported *in vivo* evidence that EPO protects neuron against ischemia-induced cell death and suggest that EPO could have neuroprotective effect by reducing the nitric oxide-mediated formation of free radicals or antagonizing their toxicity. In anemic patients the endogenous EPO is very low due to chronic renal failure. Recombinant human-EPO (rhEPO) is used in the treatment of anemia resulting from reduced production of endogenous EPO in renal failure, and in the treatment of other chronic anemias, such as those due to severe infections (Ramos *et al*, 2003).

The EPO gene was first cloned in 1985 (Jacobs *et al.*, 1985; Lin *et al.*, 1985). This sialoglycoprotein hormone consist of 165 amino acids that form a single polypeptide chain

containing two intra-chain disulfide bonds (Cys7-161 and Cys 29-33) and four potential glycosylation sites. It has three *N*-linked (Asn24, Asn38 and Asn83) and one *O*-linked (Ser126) glycosylation sites (Lin et al., 1985; Ramos et al., 2003). The molecular mass of EPO is 30-34 kDa, but the molecular mass of the peptide chain only is about 18 kDa. About 40% of its molecular weigth is due to carbohydrate moieties. Glicosylation may affect protein stability (Narhi et al., 1991). The number of sialic acids and *N*-glycosylation pattern determine the pharmacodynamic, rate of catabolism and biological activity of the EPO molecule (Fukuda *et al*., 1989; Imai *et al*., 1990; Higuchi *et al*., 1992). The protein synthesis was differed than the glycosylation process. The polypeptide synthesis was regulated genetically, whereas the oligosaccharide chains that are attached were processed by a series of enzyme reactions through post-translational modifications. Thus, a glycoprotein generally appears as a mixture of glycoforms. These glycoform populations have been shown to be cell and species as well as polypeptide and site specific. Each glycoprotein, therefore, has a reproducible and unique glycosylation profile or glycosylation pattern (Skibeli *et al*., 2001).

Recombinant hEPO has been expressed in mammalian cells, such as CHO, BHK and COS cells (Jacobs *et al*, 1985; Lin *et al*, 1985; Wen *et al*, 1994). It has also been expressed in various host, such as Tobacco plant cells (Matsumoto *et al*, 1995; Cheon *et al*, 2004), *Spodoptera frugiperda* insect cells (Wojchowski *et al*, 1987) and in yeasts such as *Saccharomyces cerevisiae* (Elliot *et al*, 1989) and *Pichia pastoris* (Burn *et al*, 2002). Every host cells has its advantages and disadvantages. However, important factor such as post-translational modifications, including protein folding and glycosylation should be carefully considered. Commercial rhEPO preparation is produced by expression of EPO cDNA clones in mammalian cells such as CHO (*chinese hamster ovary*) or BHK (*baby hamster kidney*) cells. The recombinant protein is homogenous with respect to peptide sequence of natural EPO, but heterogenous with respect to glycosilation profile which appears to differ between preparations (Lai *et al*, 1986; Skibeli *et al*, 2001). However, EPO production in mammalian cells is very expensive and production yield is not as high as in microbes. So, it is worthy to investigate if therapeutic glycoprotein such as EPO could be produced in a microorganism system such as *P. pastoris*.

In the last decade, *P. pastoris* has been used as host for various complex eukaryotic proteins expression. The yeast has a full eukaryotic system for protein processing, which is similar to mammalian, insect or plant cells. *P. pastoris* was reported to be able to express correctly human or mammalian proteins such as human-Insulin, h-IFN-β1, h-IFN-α2b, h-TNF-α, mouse-EGF, h-α-glucosidase, h-EPO mutant R103A (Kjeldsen *et al*, 1999; Skoko *et al*, 2003; Garcia *et al*, 1995; Sreekrishna *et al*, 1989; Clare *et al*, 1991; Yao *et al*, 2002; Burn *et al*, 2002) and some antibodies (scFv) (Fischer *et al*, 1999).

Different level of heterologous genes expression could be attributed by several factors such as characteristic and origin of the target proteins or genes, different codon usage or codon preference, system and vector constructions, and secretion efficiency of the recombinant protein. Previous researches reported that codon preference optimation had increased heterologous protein expression in *P. pastoris*. Such examples were the *Cyt2Aa1* toxin gene from *Bacillus thuringiensis* and a vaccine *EBA-F2* from *Plasmodium falciparum*. They were not be able to be expressed in their native sequences in *P. pastoris* due to different *codon usage* and GC-content (Gurkan and Ellar, 2003; Yadava and Ockenhouse, 2003). The genes had been successfully expressed in *P. pastoris* after their codons had been changed into the *Pichia* preference codons.

The aim of this research is to construct a synthetic version of human *EPO* gene, named *EPOsyn*, which contains *P. pastoris* codons. Codon optimation of *EPO* gene into yeast codon would hopefully increase its recombinant EPO expression compared to its native sequence in the *Pichia* expression system. Method of the synthetic gene construction and the results will be discussed.

Materials and Methods

EPO synthetic gene (EPOsyn) sequence preparation

The human EPO mRNA sequence (GenBank access no. NM_000799) or hEPO protein sequence (NP_000790) was used as reference sequence to design the *EPO*-synthetic

gene (*EPOsyn*) sequence (Table 1). In *EPOsyn* sequence, a number of codons have been changed into yeast preference codons of *P. pastoris*. The yeast preference codons were retrieved from a codon usage database (http://www.kuzusa.or.jp/codon). The open reading frame (ORF) sequence of *EPO* gene includes 501 nucleotides coding for 166 amino acids (without the signal sequence). Each of amino acids codons of the native hEPO were changed into the most preferred codon of *P. pastoris* (Table 2). Most of those codon changes involved the last base of the triplet. All the changes of the codons were resumed in Table 3.

The synthetic gene sequence was divided into 8 parts, in which 8 single stranded oligonucleotides were synthesized and ordered to construct the gene (Promega). Each oligonucleoties, named P1, P2, P3, P4, P5, P6, P7 and P8, have an average length of 90 nucleotides (nt), except the P8 (66 nt). All oligos have 27 to 30 nt overlap regions with the adjacent oligos at both 3'- and 5'-ends to facilitate ligation between each oligos. A pair of short primers (30 nt) was prepared for gene amplification. Those primers, FP1 and RP8, contain *Xho*I and *Xba*I (or *Sal*I) restriction sites, respectively. Another pair of primers, RP4 and FP5, were also prepared for partial gene amplification (See Figure 1).

Construction method of the EPO-synthetic gene (EPOsyn)

The synthetic gene (EPOsyn) was constructed by a *recursive*-PCR (rPCR) method, a PCR-based gene synthesis methods (Casimiro *et al.*, 1997). A number of 8 oligos (P1 ~ P8) with an average length of 90 nt were mixed into a PCR reaction to yield a full length synthetic gene. The oligos have partially overlapping region between the adjacent oligos. The full length gene construction was divided into two parts. Thus the 8 oligos were divided into 2 groups, in which each group contains 4 oligos (P1~P4 and P5~P8). This procedure was carried out to avoid undesired pairing of oligos and reduces number of DNA contaminants in the PCR product. The full length gene synthesis was completed in two steps PCR reactions. The first step is the formation of half of the gene sequence and the second step is the formation of the full length gene (Figure 1B). Two pairs of short primers (FP1/RP4 and FP5/RP8) were used to amplify the gene sequence in the first step PCR. *Pfu* DNA polymerase (Fermentas) was used during the synthetic gene construction. PCR reactions of the first step PCR were carried out in the following cycles: first denaturation (95°C, 2min), denaturation (95°C, 1min), annealing (50°C, 1 min), extension (72°C, 2min) in 30 cycles, and final extension (72°C, 5min). The PCR products were analysed in agarose gel electrophoresis and the DNA bands of expected length were isolated and purified.

Table 1. Mature protein and mRNA sequences of the human *EPO* used as the reference sequences in designing the *EPO*-synthetic gene

Access. no.	Sequences
NP_000790	APPRLICDSR VLERYLLEAK EAENITTGCA EHCSLNENIT VPDTKVNFYA WKRMEVGQQA VEVWQGLALL SEAVLRGQAL LVNSSQPWEP LQLHVDKAVS GLRSLTTLLR ALGAQKEAIS PPDAASAAPL RTITADTFRK LFRVYSNFLR GKLKLYTGEA CRTGDR*
NM_000799	gccccacc acgcctcatc tgtgacagcc gagtcctgga gaggtacctc ttggaggcca aggaggccga gaatatcacg acgggctgtg ctgaacactg cagcttgaat gagaatatca ctgtcccaga caccaaagtt aatttctatg cctggaagag gatggaggtc gggcagcagg ccgtagaagt ctggcagggc ctggccctgc tgtcggaagc tgtcctgcgg ggccaggccc tgttggtcaa ctcttcccag ccgtgggagc ccctgcagct gcatgtggat aaagccgtca gtggccttcg cagcctcacc actctgcttc gggctctggg agcccagaag gaagccatct cccctccaga tgcggcctca gctgctccac tccgaacaat cactgctgac actttccgca aactcttccg agtctactcc aatttcctcc ggggaaagct gaagctgtac acaggggagg cctgcaggac aggggacaga tga

Note: The leader sequence (first 27 amino acids) is not shown. '*' is stop codon (tga)

The first PCR produced two part of partial target gene. The second step PCR reaction was carried out to ligate both parts of gene and produce the full-length target gene (Figure 1B). PCR conditions were the same as the first step PCR, except that the annealing temperature was 53°C, and a pair of terminal primers (FP1 and RP8) was added into the PCR mixture. The PCR cycles are shown in Figure 3B and 3D. The DNA band of around 520 bp was isolated and purified from agarose gel with a DNA column purification kit.

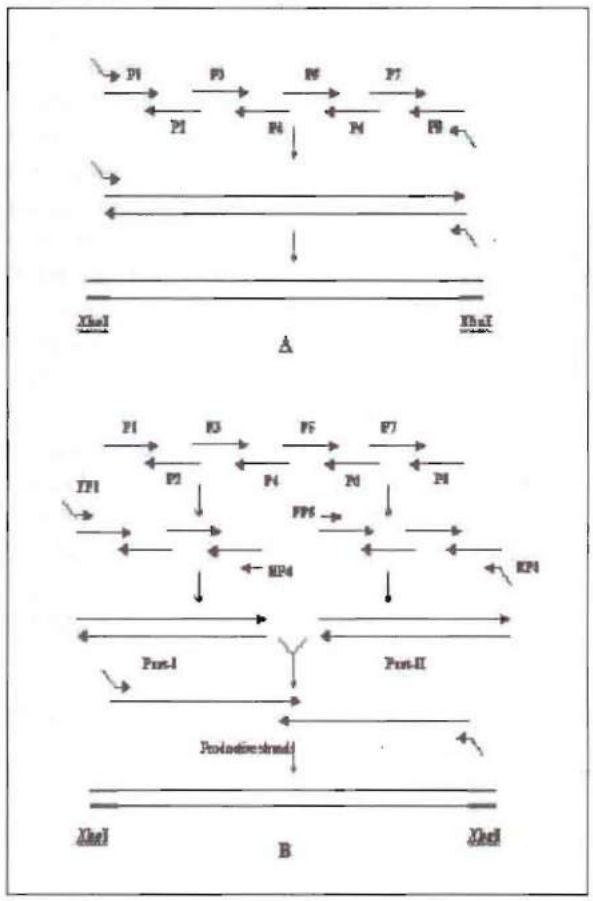

Figure 1. A scheme of *recursive*-PCR methods used in constructing the *EPO*- synthetic gene (*EPOsyn*). (**A**) Method-1: all of oligos were mixed in a single PCR tube and were reacted in a single PCR reaction. The PCR product was then amplified with a pair of short primers. (**B**) Method-2: all of oligos were divided into two parts, each different PCR mixtures contains 4 oligos. Then both PCR products were amplify and extended by a second PCR reaction to form the full-length gene sequence using a pair of terminal primers (FP1 and RP8).

Cloning and amplification of the EPO-synthetic gene

The full-length *EPO*-synthetic gene (*EPOsyn*), that had been column purified, was incubated with *Taq* DNA polymerase and dNTPs at 72°C for 30 minutes in order to give an A-overhang at both ends of the gene for cloning purpose. The gene was again column purified and cloned into pCR2.1 cloning vector (Invitrogen), resulting pCR-EPS recombinant plasmids. Those plasmids were transformed into *E. coli* TOP10F'. Screening of recombinant plasmids was carried out by restriction analysis to find out the correct recombinant having the right insert DNA.

Transformation, DNA extraction and restriction analysis were carried out according to basic molecular biology protocols (Ausubel *et al.*, 2002).

Restriction analysis and gene sequencing

Screening of the recombinant pCR-EPS plasmids was carried out in two steps. First, plasmids were restricted with *Eco*RI, which was part of the MCS sites of the plasmid, which would release the insert DNA from the correct recombinant plasmids. The positive clone(s) would have a DNA insert of around 520 bp. Second step, the positive clone(s) were again restricted with double restriction enzymes, *Xho*I and *Xba*I. Both restriction sites were incorporated into the *EPOsyn* sequence at 5'-end and 3'-end of the gene respectively. The positive clone(s) which showed the insert DNA (520 bp) were isolated and purified prior to DNA sequencing analysis. DNA isolation and extraction from agarose gel were done using DNA extraction kit and carried out according to the manufacturer's protocols (Roche).

Results and Discussions

The EPO-synthetic gene design and codon optimation

The EPO-synthetic gene, named *EPOsyn*, had been designed to contain the most preferred codon of *P. pastoris* by changing human codon from its native *EPO* sequence into yeast codon. Despite any computer programs that were recently available for designing a synthetic gene sequence, we have designed the *EPO*-synthetic gene (*EPOsyn*) sequence manually. The choice of codon usage that was used in the gene sequence was determined empirically by information from

an available codon usage database (www.kazusa.or.jp/codon). The aim of synthetic gene construction is, among others, to facilitate transcription and translation processes in a specific host cell, thus to optimize the gene expression in a given host cell. This could be achieved by changing the amino acid codon of the original gene into the preference codon of the host. In this work, we have designed a version of *EPO*-synthetic gene (*EPOsyn*) sequence that contains *P. pastoris*'s codon preference instead of human codon. The codon change involved almost all of amino acids present in the human EPO, except for methionine and thryptophane. Table 2 and Table 3 showed codon distribution and codon usage of *P. pastoris* and that of human respectively obtained from codon usage database (www.kazusa.or.jp/codon). The tables show that some amino acids use certain codons more often than others for the same amino acid. Yeast codon usage is significantly different from the human one as shown on those tables.

The mature human EPO contains 166 amino acids (Lin *et al*, 1985; Jacobs *et al*, 1985). There are 20 types of amino acid found in hEPO, which are coded by 51 different codons (Table 4). The gene has 59.08% GC content, with signal sequence and stop codon excluded. Almost all of amino acid codon in the native EPO gene had been changed in the synthetic gene construct (*EPOsyn*), except those of methionine (ATG) and tryptophane (TGG). The synthetic gene sequence designed in this work, contains one most preferred codon for each amino acid. Thus, there is one codon only for one type of amino acid, except for Ala. There is one codon (GCC) used for one residue of Ala out of 19 residues (GCT). This change was done inevitably to avoid undesirable restriction site within our gene. The most preferred yeast codon was determined from the *P. pastoris* codon database (Table 2). Therefore, there are 20 codons that code for 20 different amino acids used in the synthetic gene (*EPOsyn*), compared to 51 codons used in the native gene (*EPO*).

Tabel 2. Codon usage in *Pichia pastoris* compiled from 85 CDS datas (around 49099 codons)

AA	Codon	Frequence (/1000)	Fraction	AA	Codon	Frequence (/1000)	Fraction
G	GGG	6.01	0.10	M	**ATG**	19.31	**1.00**
	GGA	19.49	0.32	I	ATA	11.26	0.18
	GGT	27.62	**0.45**		**ATT**	31.37	**0.50**
	GGC	8.21	0.13		ATC	19.96	0.32
E	GAG	30.18	0.44	T	ACG	5.97	0.11
	GAA	38.53	**0.56**		ACA	13.56	0.24
D	**GAT**	36.01	**0.58**		**ACT**	22.75	**0.40**
	GAC	26.80	0.42		ACC	14.40	0.25
V	GTG	12.26	0.19	W	**TGG**	10.10	**1.00**
	GTA	9.63	0.15	STOP	TGA	0.31	0.18
	GTT	27.50	**0.42**	C	**TGT**	8.25	**0.66**
	GTC	15.56	0.24		TGC	4.28	0.34
A	GCG	3.60	0.05	STOP	TAG	0.49	0.29
	GCA	15.23	0.23	STOP	**TAA**	0.92	**0.54**
	GCT	30.00	**0.45**	Y	TAT	14.42	0.43
	GCC	17.33	0.26		**TAC**	18.96	**0.57**
R	AGG	6.44	0.15	L	**TTG**	32.08	**0.33**
	AGA	20.55	**0.48**		TTA	14.83	0.16
	CGG	2.02	0.05		CTG	15.28	0.16
	CGA	4.44	0.10		CTA	10.79	0.11
	CGT	6.92	0.16		CTT	15.64	0.16
	CGC	2.18	0.05		CTC	7.33	0.08
S	AGT	11.98	0.15	F	**TTT**	23.67	**0.54**
	AGC	7.17	0.09		TTC	20.51	0.46
	TCG	6.76	0.08	Q	CAG	14.70	0.38
	TCA	14.91	0.18		**CAA**	23.91	**0.62**
	TCT	23.50	**0.29**	H	**CAT**	11.00	**0.53**
	TCC	16.29	0.20		CAC	9.63	0.47
K	**AAG**	36.01	**0.55**	P	CCG	3.93	0.09
	AAA	29.43	0.45		**CCA**	17.88	**0.41**
N	AAT	22.51	0.46		CCT	15.21	0.35
	AAC	25.91	**0.54**		CCC	6.46	0.15

Source: Codon data base (http://ww.kuzusa.or.jp/codon/); The most preferred codon is in bold.

However, when we analyzed the codon usage of the *alcohol oxidase* gene, which is the highly expressed gene in *P. pastoris* when grown in methanol, we found that some amino acids preferred a rather different codon. Those amino acids and its codon and fraction are Glu (GAG, 0.53), Asp (GAC, 0.74), Ser (TCC, 0.45), Phe (TTC, 0.81) and His (CAC, 0.83). Therefore, we preferred to use those codons (GAG, GAC, TCC, TTC and CAC) for those 5 amino acids (Glu, Asp, Ser, Phe and His) respectively in our synthetic gene. The codon and its number of occurrence within the gene sequence are shown in Table 4. Table 4 also showed codon distribution in the human *EPO* gene (*wildtype*), as well as in the 'yeast' *EPO* gene (*EPOsyn*).

When the codons were changed then naturally the gene sequence was also changed. Consequently, the GC content of the synthetic gene was also changed accordingly. It was known that different organisms have different GC contents. It is a specific character of each organism. Human genes have around 60% of GC, based on Codon usage database (www.kazusa.or.jp/codon/). On the other hand, yeast genomes contain unfortunately less GC than the human one. It has around 43% to 47% GC content. The difference is quiet significant that may affect heterologous human gene(s) expression in yeast. In this version of *EPOsyn* gene, we have changed the GC content from 59.08% in native gene to 45.31% in synthetic one (Table 5).

This value is approaching the real GC content of yeast native genes, so that a better expression is expected to occur and no hurdle in gene expression is going to be faced due to different GC content. Although the number of codons used decreased from 51 to 20 types of codon (excluding Stop codon) (Table 5), the synthetic gene would benefit the abundance of the tRNAs of preferable codon usage (Ikemura, 1981).

Table 4. Codon distribution of different type of amino acids within the native (*wildtype*) and synthetic *EPO* genes

Amino Acid	Codon	Number of AA / Codon		Amino Acid	Codon	Number of AA / Codon	
		EPO	*EPOsyn*			*EPO*	*EPOsyn*
G	GGG	3	-	M	ATG	1	1
	GGA	1	-	I	ATA	-	-
	GGT	-	8		ATT	-	5
	GGC	4	-		ATC	5	-
E	GAG	8	12	T	ACG	2	-
	GAA	4	-		ACA	3	-
D	GAT	2	-		ACT	4	11
	GAC	4	6		ACC	2	-
V	GTG	1	-	W	TGG	3	3
	GTA	1	-	STOP	TGA	1	-
	GTT	1	11	C	TGT	2	4
	GTC	8	-		TGC	2	-
A(*)	GCG	1	-	STOP	TAG	-	-
	GCA	-	-	STOP	TAA	-	1
	GCT	6	18	Y	TAT	1	-
	GCC	12	1		TAC	3	4
R	AGG	2	-	L	TTG	3	23
	AGA	2	14		TTA	-	-
	CGG	3	-		CTG	13	-
	CGA	4	-		CTA	-	-
	CGT	-	-		CTT	2	-
	CGC	3	-		CTC	5	-
S	AGT	1	-	F	TTT	-	-
	AGC	3	-		TTC	4	4
	TCG	1	-	Q	CAG	7	-
	TCA	1	-		CAA	-	7
	TCT	-	-	H	CAT	1	-
	TCC	3	10		CAC	1	2
K	AAG	5	8	P	CCG	1	-
	AAA	3	-		CCA	5	8
N	AAT	5	-		CCT	1	-
	AAC	1	6		CCC	1	-

(*) One residue of Ala (A) use a different codon (GCC) than the others (GCT).

There is an opinion that 'one amino acid – one codon' approach could have several drawbacks, such as the transcribed mRNA would resulting in an imbalance tRNA pool, risk of instability of the mRNA sequence and difficulties to avoid undesirable restriction sites (Gustaffsson *et al*, 2004). However, Klasen *et al* (2004) have reported that silent mutation in 'one amino acid – one codon' optimized gene can increased protein expression four-fold.

Construction of the synthetic double stranded DNA of the EPOsyn gene

The synthetic gene was constructed in a two-step PCR method. First step is partial gene construction of the target gene from two groups of 4 oligos (P1~P4 and P5~P8). Each group of oligos formed half of the gene sequence. The target DNA fragment was isolated and purified from the agarose gel. In the second step, both parts of the target gene were assembled into a full-length gene sequence using a PCR-based method. Figure 2 shows rPCR products of 4 long oligonucleotides assembly, each represent half of the gene sequence (around 270 bp). The PCR products were not as clean as expected since there were also 'unproductive' strands that would not amplify to produces the target DNA sequence. The 'productive' strands only (which have the overlap region at the 3'-end) that were amplified to produce the target gene sequence (Figure 2).

Both partial gene sequences (~270 bp) were isolated and purified from the gel, named part-I and part-II as in Figure 1. Both parts of the synthetic gene were then assembled in a second step PCR reaction to form the full-length synthetic gene. The full-length gene assembly was carried out in two ways as shown in Figure 3. First, both parts of the synthetic gene were mixed with a pair of short primer (FP1 and RP8) in a PCR tube. After 30 cycles of reactions, the PCR product was analyzed on agarose gel (Figure 3A, 3B). Second, both parts of the synthetic gene were mixed and initially amplified for 10 cycles, enough to form full-length DNA templates. It was

Tabel 5. Number of different amino acids, codon used, nucleotides and GC content of native gene (*EPO*) and synthetic gene (*EPOsyn*) of human EPO

Gene	Number of AA types	Number of codons used[1)]	No. of nucleotides (A+T+G+C) [2)]	Number of (G+C) [2)]	%GC [2)]
EPO	20	51	501	296	59.08
EPOsyn	20	20	501	227	45.31

[1)] Excluding Stop codon; [2)] Without leader sequence.

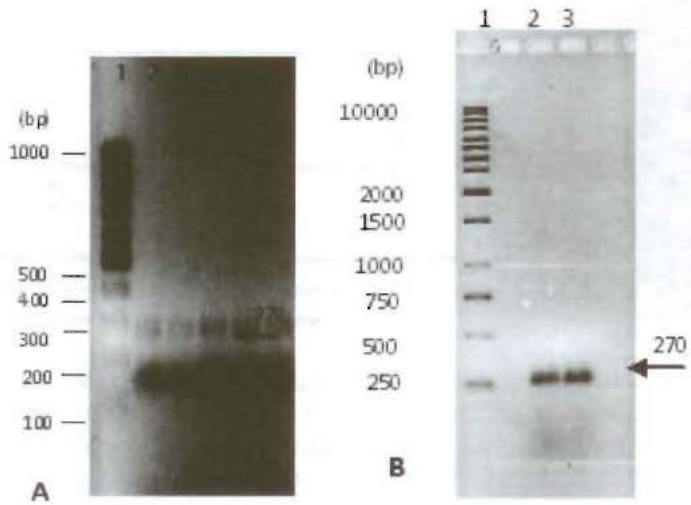

Figure 2. Construction of partial synthetic *EPO*-gene from 4 oligos that form two halves sequence of the target gene. The first part of *EPOsyn* gene made from oligos no. P1 to P4 (**A**) and the second part made of oligos no. P5 to P8 (**B**).

then amplified for the next 25 cycles (Figure 3C, 3D) using the amplifying primers (FP1 and RP8).

As shown in Figure 3, both methods produced the target DNA band of around 520 bp. The sequential PCR in the second method produced more target DNA than the first one. However, both PCR products still contain much DNA contaminants of 'unproductive' strands, since ligation occurred only on the 'productive' strands. To exclude any DNA contaminant, the full-length target gene (520 bp) was isolated and purified from the gel.

Gene cloning and restriction analysis of the recombinant plasmid

The purified full-length synthetic gene was used for gene cloning. After incubation with a *Taq* DNA polymerase to provide an A-overhang and subsequent column purification, the synthetic gene was then cloned into pCR2.1 cloning plasmid (Invitrogen) and transformed into *E. coli* XL-1 Blue. The resulted plasmid pCR-EPOsyn was screened by restriction analyses. First screening was done by a single digestion with *Eco*RI, which would release the DNA insert (Figure 4A). The *Eco*RI sites were found on both sites of the MCS region of the plasmid, flanking the insert DNA

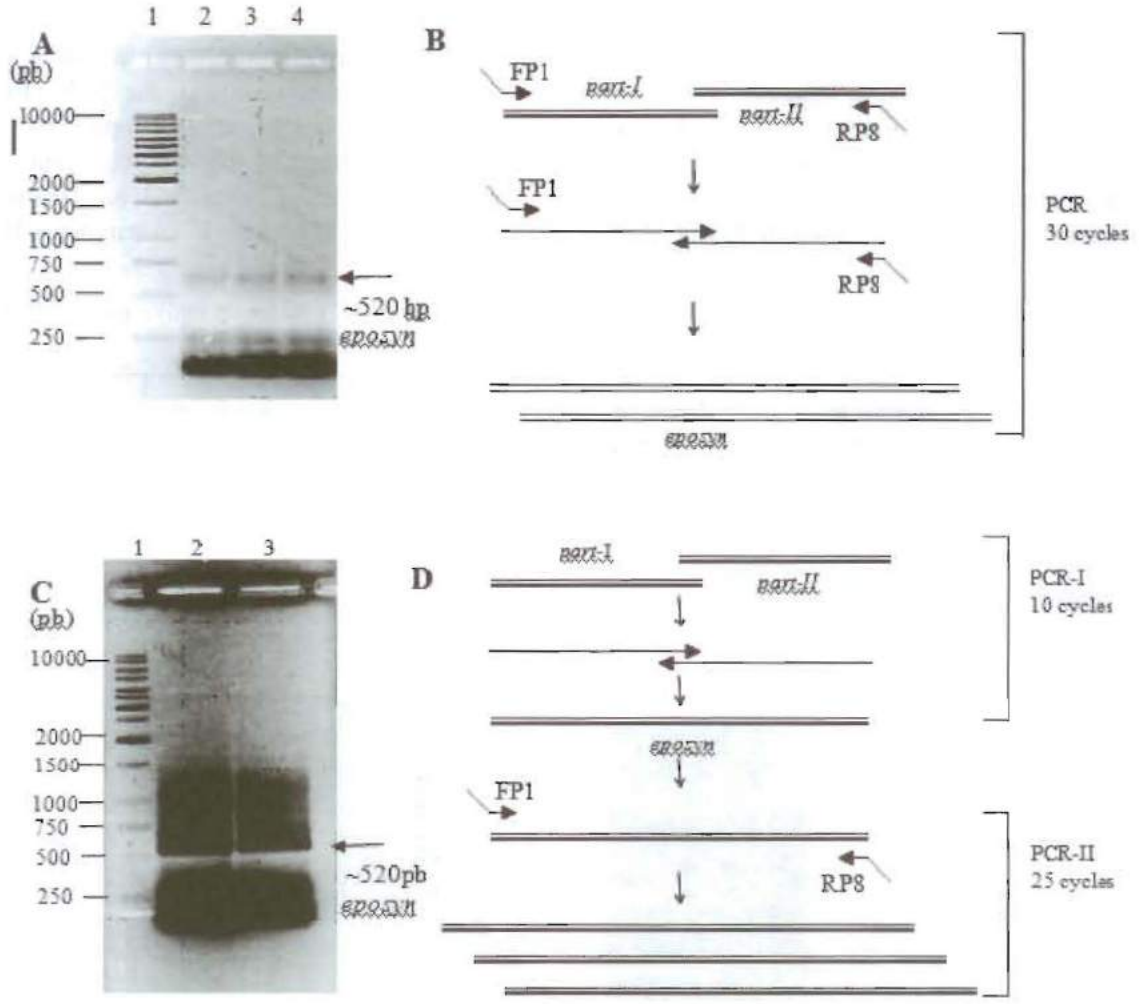

Figure 3. Full-length *EPOsyn* gene synthesis from two partial gene sequences in a single PCR reaction (A) and two successive PCR reactions (C). First method, both parts of the synthetic gene (*part-I* and *part-II*) were mixed in a single PCR reaction (30 cycles) with addition of terminal primers (FP1 and RP8) (A,B). Whereas in the second method, both parts of the gene were firstly placed in a PCR reaction (10 cycles), then successively placed in the second PCR reaction (25 cycles) with the addition of terminal primers (C,D). The target product is a DNA band of around 520bp in length as is shown in on the left figures (A,C). Schematic process of the full-length gene amplification of *EPO*-synthetic is shown on the right figures (B,D)

sequence. The positive clones were then analyzed with the second screening using *Xho*I and *Xba*I double digestions. Both restriction sites were incorporated into the gene sequence at the 5'-end and 3'end respectively.

Restriction analysis of recombinant plasmids resulted in some positive clones. Those clones positively showed a DNA insert of 520 bp after digestions with both *Xho*I and *Xba*I, as well as with *Eco*RI (Figure 4B). Figure 4B showed some positive clones digested with both *Xho*I and *Xba*I. A number of 8 positive clones out of 70 analyzed were obtained.

DNA sequence analyses of the synthetic EPO-gene

A number of 8 independent clones were sequenced to determine the error rate of the gene synthesized. The cloned DNA showed an error rate of 1.59% from a total of 3984 base pairs sequenced from the 8 clones. There were 2 types of mutation observed, deletion and substitution. While 1.42% of error rate was due to deletions, 0.17% of error was due to substitutions. The mutations were located at different sites, spread over the target gene that was synthesized. Mutation types and locations are summarized in Table 6. Figure 5 shows DNA sequences of 8 independent clones and the target gene (*EPOsyn*).

The synthesis of double stranded DNA in this research used PCR product of previous assembly cycle as primers for the next assembly cycle (Figure 1). Therefore the fidelity of the PCR reaction during gene assembly becomes an important factor that influences the accuracy of the target DNA sequence. Any error in previous assembly would be progressively accumulated in the next step. Polymerase fidelity is influenced by some factors such as the tendency of the enzyme to include an incorrect nucleotide, the presence of a proof reading 3' to 5' exonuclease activity and the ability to extend mismatches (Cline *et al*, 1996; Singh *et al*, 1996). Therefore we used an enzyme with high fidelity, such as *Pfu*, which is a thermostable DNA polymerase derived from *Pyrococcus furiosus* that showed those promises. The error rate of *Pfu* polymerase was reported to be 5-fold lower than that of *Deep Vent* polymerase and 9 to 10-fold lower than *Taq* polymerase (Flaman *et al*, 1994). Cline *et al* (1996) has reported a more detailed PCR fidelity of some DNA polymerases, in which the average error rate increased as follows: *Pfu* ($1.3x10^{-6}$) < *Deep Vent* ($2.7x10^{-6}$) < *Vent* ($2.8x10^{-6}$) < *Taq* ($8.0x10^{-6}$) << exo⁻ *Pfu* and *UlTma* ($\sim 5x10^{-5}$). In exception of *Taq* polymerase, all other polymerases mentioned above are proofreading enzymes. It is clear that *Pfu* is among the best DNA polymerase having high replication fidelity.

However, in the case of synthetic gene assembly, any mutations or errors in the DNA sequence was not contributed by the DNA polymerase performance only. It is also influenced by other factors, such as the homogeneity and accuracy of the synthetic oligonucleotides used, method of gene assembly being used, and various factors involved in the PCR reactions in which the synthetic gene was assembled. The length of oligonucleotides used in the gene assembly might also contribute to the accuracy of successfully constructed gene sequence.

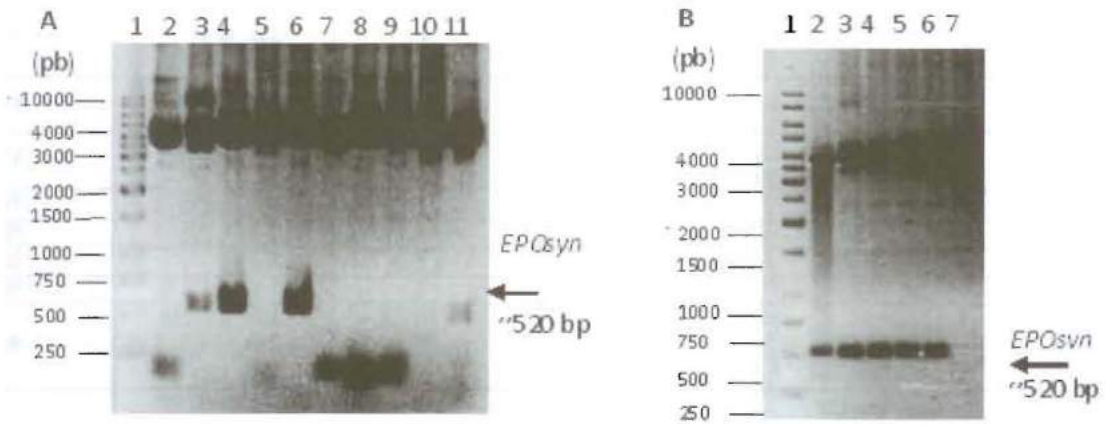

Figure 4. Restriction analyses of the recombinant plasmid pCR-*EPOsyn* digested with *Eco*RI (**A**) and with double enzymes *Xho*I and *Xba*I (**B**). The positive clones have a DNA insert of around 520 bp

Singh *et al.* (1996) has reported a construction of a 701 bp synthetic gene with a comparable gene construction strategy used in this research, except that they used *Deep Vent* polymerase. They have an error rate of 0.68%, in which 0.61% are substitutions and 0.07% are deletions. Whereas Young and Dong (2004) reported a construction of three genes (A1, A2, A3) having 1.092 to 1.195nt in length using the same *Pfu* polymerase. They ended with 1.21% of error rate, in which all of those errors were deletions. Whereas in our gene (501 nt) there are more deletions (1.42%) than substitutions (0.17%) (Table 6).

The error rate observed in our research (1.59%) was slightly higher than that was reported previously : 0.68% (Singh et al, 1996), 1.21% (Young and Dong, 2004), 0.24% (Chen, 1994), and 0.32% (McLain, 1986). Although the error rate seems to be higher than that was reported by others, the gene synthesis approach in this research is somehow different. We used a small number of oligonucleotides (8 oligos) having a quiet long nucleotides (90 nt in length in average) with 20 to 30 nt overlap between oligos and *Pfu* polymerase. Singh *et al* (1996) used a large number of oligos (28 oligos) having 40 to 45 nt in length and 14 to 18 nt overlap between the adjacent oligos and *Deep Vent* polymerase. Whereas Young and Dong (2004) used 28 to 32 oligonucleotides of 50 nt in length with 10 nt overlap between adjacent oligos and *Pfu* polymerase. Compared to short oligos (40 or 50 nt in length), long oligonucleotides having more than 80 nt in length need to be carrefully prepared. Though all of oligo primers were column purified, any error in oligos sequence could not be easily detected in the oligo preparations. This is one of disadvantages when using very long oligos. However, using a small number of long oligos in the synthetic gene construction might present other benefits such as low cost of the oligonucleotides prepartion, and the target gene would be efficiently assembled with less number of oligo.

Among 8 different clones which had been sequenced, we have 4 best clones having 2 or 3 point mutations (Table 6). One of theme, clones no. C55 or G3, was used to correct the synthetic gene sequence by site directed mutagenesis to obtain the desired corrected sequence. The corrected gene was again cloned into pCR2.1 cloning plasmid and sequenced. The resulting plasmid pCR-EPS was used for subsequent cloning into pPICZα expression vector for transformation and protein expression in *P. pastoris*.

Table 6. Number, type of mutation(s) and sequence similarity from 8 independent clones of *EPO*-synthetic gene produced by rPCR-based method

No	Clone No.	Type of mutation	Number of mutation (nt)	Mutation position (nt no.)	Sequence similarty (%)
1	C17	Deletion Substitution	11 1	22, 345-352, 363, 424. 125	97.6
2	C27a	Deletion Substitution	10 3	406, 423-431. 92, 110, 442.	97.4
3	C27b	Deletion Substitution	10 3	406, 423-431. 92, 110, 442.	97.4
4	C55	Deletion	3	4, 294, 295	99.4
5	G3	Deletion	3	4, 294, 295	99.4
6	H10	Deletion	2	4, 18.	99.6
7	K2	Deletion	3	190, 191, 278.	99.4
8	L5	Deletion	17	100, 237-248, 276, 309, 319, 328	96.6
Total	Deletion Substitution	59 7	1.42 % 0.17 %		

*) nt : nucleotide.

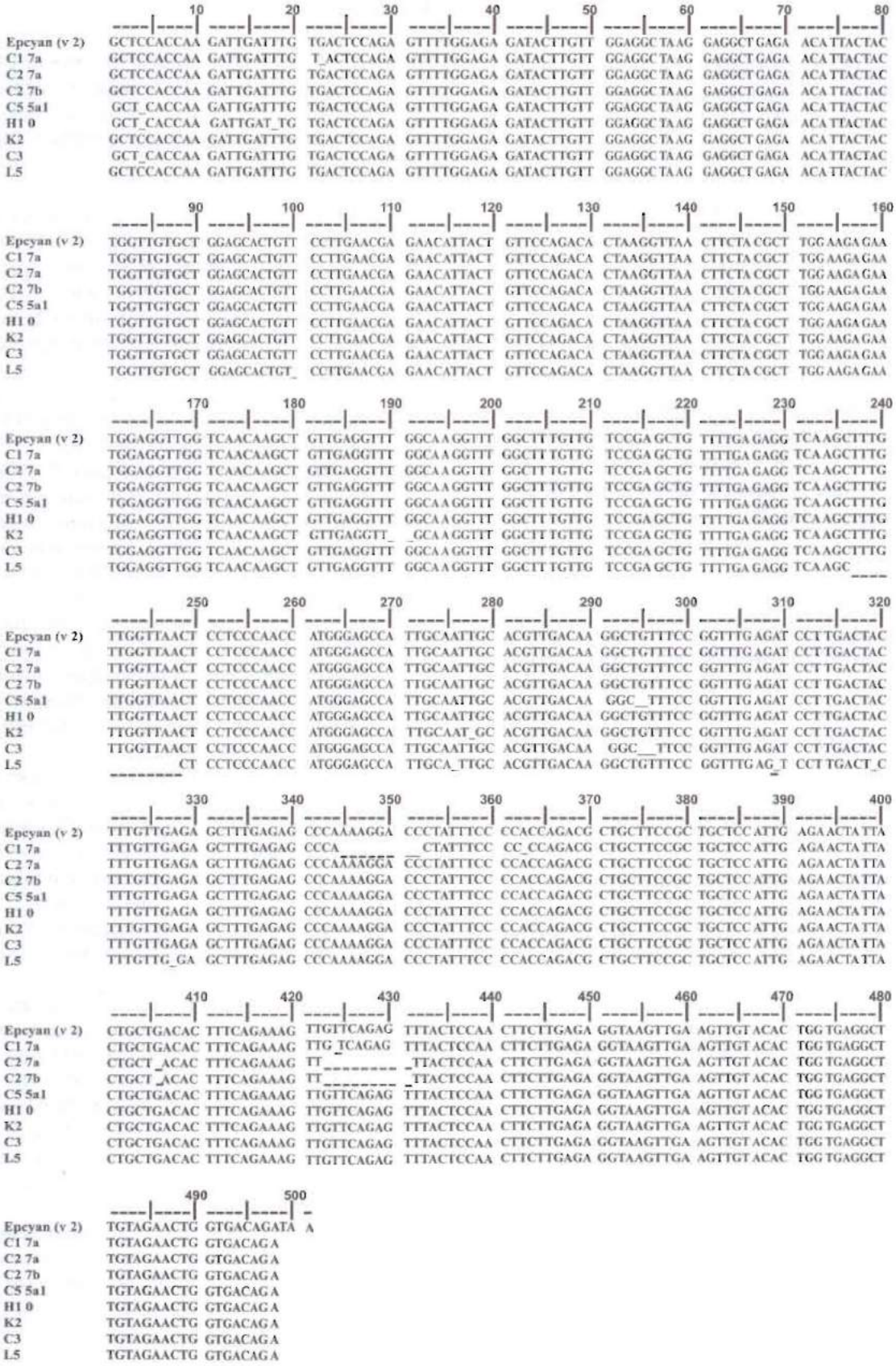

Figure 5. DNA sequence analysis of 8 independent clones of *EPO*-synthetic gene produced. The first row is the target synthetic gene sequence (*EPOsyn*)

Conclusions

An EPO-synthetic gene had been successfully constructed using a recursive-PCR (rPCR) method from 8 synthetic single strand oligonucleotides with an average length of 90 nt. The synthetic gene has 520 base pairs in length with optimized *P. pastoris* codon. Each amino acid present in the synthetic gene was represented by one most preferred codon only. There are 20 codon used in the synthetic gene compared to 51 codon present in the native (human) gene. The GC content had been decreased from 59.08% in the native gene to 45.31% in the synthetic gene, approaching the natural GC content of *P. pastoris* (43~47%). Gene synthesis in a single mixture of 8 oligos in a PCR reaction at once was not successful, due probably to un-optimized oligos concentration and PCR conditions. However, synthesis using a two-step PCR method by partial gene construction has successfully resulted in a full-length gene sequence. Cloning of the full-length synthetic gene into pCR2.1 plasmid cloning has successfully obtained several positive clones containing the 520 bp *EPOsyn* gene. DNA sequence analysis has revealed the *EPOsyn* gene sequence that had been produced. The synthesis method used in this research has an error rate of 1.59%, in which 1.42% of this error rate was due to deletions and 0.17% was due to substitutions. None of the 8 independent clones has perfectly matched target sequence. However, several clones have 2 point mutations only, which would be easily corrected by site directed mutagenesis. Four best clones (C55, G3, H10 and K2) have sequence similarity of 99.4 to 99.6%. One of them (C55 or G3) had been used in the next experiment to correct the DNA sequence. A clone with perfectly matched sequence of the target gene had been obtained.

Acknowledgment

We thank our collaborator in the EPO project Dr. Adi Santoso for his endless support and supply of some materials. We also thank Yuliawati, Dian Fitria Agustiyanti and Aminah for their valuable assistances. We thank the Research Center for Biotechnology - Indonesian Institute of Sciences for providing funds.

References

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA & Struhl K (editors) (2002) Short Protocols in Molecular Biology 5th edition: A Compendium of Methods from Current Protocols in Molecular Biology. John Wiley & Sons, Inc. Canada.

Burns S, Arcasoy MO, Kurian E, Selander K, Emanuel PD and Harris KW (2002) Purification and characterization of the yeast-expressed erythropoietin mutant Epo (R103A), a specific inhibitor of human primary hematopoietic cell erythropoiesis. Blood 99 (12): 4400-4405.

Casimiro DR, Wright PE and Dyson HJ (1997) PCR-based gene synthesis and protein NMR spectroscopy. Structure 5:1407-1412.

Chen GQ, Ramachandran B and Gouaux JE (1994) Total gene synthesis: novel sinle-sep and convergent strategies applied to the construction of a 779 bp bacteriorhodopsis gene. J Am Chem Soc 116: 8799-8800.

Cheon BY, Kim HJ, Oh KH, Bahn SC, Ahn JH, Choi JW, Ok SH, Bae JM and Shin JS (2004) Overexpression of human erythropoietin (EPO) affects plant morphologies: retarded vegetative growth in tobacco and male sterility in tobacco and *Arabidopsis*. Transgenic Research 13: 541-549.

Clare JJ, Romanos MA, Rayment FB, Rowedder JE, Smith MA, Payne MM, Sreekrishna K, and Henwood CA (1991) Production of Epidermal Growth Factor in yeast: High-level secretion using *Pichia pastoris* strains containing multiple gene copies. Gene 105: 205-212.

Ebert BL and Bunn HF (1999) Regulation of the Erythropoietin Gene. Blood 94(6): 1864-1877.

Elliott S, Griffin J, Suggs S, Lau EP and Banks AR (1989) Secretion of glycosylated human erythropoietin from yeast directed by the α-factor leader region. Gene 79: 167-180.

Elliott S, Lorenzini T, Asher S, Aoki K, Brankow D *et al* (2003) Enhancement of therapeutic protein *in vivo* activities through glycoengineering. Nature Biotechnology 21: 414-421.

Fischer R, Drossard J, Emans N, Commandeur U and Hellwig S (1999) Toward molecular farming in the future: *Pichia pastoris*-based production of single-chain antibody fragments. Biotechnol Appl Biochem 30: 117-120.

Fukuda MN, Sasaki H, Lopez L, and Fukuda M (1989) Survival of recombinant erythropoietin in the circulation: The role of carbohydartes. Blood, 73: 84-89.

Garcia JN, Aguiar JA, Gill M, Alvarez A, Morales J, Ferrero J, Gonzales B, Padron G, and Menendez A (1995) High level expression of human IFN-α2b in *Pichia pastoris*. Biotechnologia Aplicada 12: 152-155.

Gerngross TU (2004) Advances in the production of human therapeutic proteins in yeasts and filamentous fungi. Nature Biotechnology 22 (11): 1409-1414.

Gurkan C and Ellar DJ (2003) Expression of the Bacillus thuringiensis *Cyt2Aa1* toxin in *Pichia pastoris* using a synthetic gene construct. Biotechnol Appl Biochem 38: 25-33.

Gustafsson C, Govindarajan S and Minshull J (2004) Codon bias and heterologous protein expression. Trend in Biotech Vol. 22 : 346-355.

Higuchi M, Oh-Eda M, Kuboniwa H, Tomonoh K, Shimonaka Y and Ochi N (1992) Role of sugar chains in the expression of the biological activity of human erythropoietin. Journal of Biological Chemistry 267 (11): 7703-7709.

Ikemura T (1981) Correlation between the abundance of *E. coli* transfer RNAs and the occurance of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. Journal of Molecular Biology 151 (3): 389-409.

Imai N, Kawamura A, Higuchi M, Oh-Eda M, Orita T, Kawaguchi T, and Ochi N. Physicochemical and biological comparison of recombinant human erythropoietin with human urinary erythropoietin. Journal of Biochemistry, 107: 352-359.

Jacobs K, Shoemaker C, Rudersdorf R, Neil SD, Kaufman RJ, Mufson A et al (1985) Isolation and characterization of genomic and cDNA clones of human erythropoietin. Nature 313: 806-810.

Kjeldsen T, Pettersson AF and Hach M (1999) Secretory expression and characterization of insulin in *Pichia pastoris*. Biotechnol Appl Biochem 29: 79-86.

Klasen M and Wabi M (2004) Silent point mutation in DsRed resulting in enhanced relative fluorescence intensity. BioTechniques 36 (2): 236-237.

Koury MJ and Bondurant MC (1992) The molecular mechanisme of erythropoietin action. European Journal of Biochemistry 210: 1-14.

Lai PH, Everett R, Wang FF, et al (1986) Structural characterization of human erythropoietin. Journal of Biological Chemistry 261: 3116-3121.

Lin FK, Suggs S, Lin CH, Browne JK, Smalling R, Egrie JC *et al* (1985) Cloning and expression of the human erythropoietin gene. PNAS USA 82: 7580-7584.

Matsumoto S, Ikura K, Ueda M and Sasaki R (1995) Characterization of a human glycoprotein (erythropoietin) produced in cultured tobacco cells. Plant Mol Biol. 27: 1163-1172.

McLain WH, Foss K, Mittelstadt KL and Schneider J (1986) Variants in clones of gene-machine-synthesized oligo deoxynucleotides. Nucleic Acids Res. 22: 3259-3260.

Ramos AS, Schmidt CA, Andrade SS, Fronza M, Raffery B and Dalmora SL (2003) Biological evaluation of recombinant human erythropoietin in pharmaceutical products. Brazilian Journal of Medical and Biological Research 36: 1561-1569.

Singh PK, Sarangi BK and Tuli R (1996) A facile method for the construction of synthetic genes. Journal of Bioscience Vol 21(6):735-741.

Skibeli V, Nissen-Lie G and Torjesen P (2001) Sugar profiling proves that human serum erythropoietin differs from recombinant human erythropoietin. Blood 98(13): 3626-3634.

Skoko N, Argamante B, Grujicic NK, Tisminetzky SG, Glisin V and Ljubijankic G (2003) Expression and characterization of human interferon-β1 in the methylotrophic yeast *Pichia pastoris*. Biotechnol Appl Biochem 38: 257-265.

Sreekrishna K, Potenz RHB, Cruze JA, McCombie WR, Parker KA, Nelles L, Mazzaferro PK, Holden KA, Harrison RG, Wood PJ, Phelps DA, Hubbard CE, and Fuke M (1988) High level expression of heterologous proteins in methylotrophic yeast Pichia pastoris. Journal of Basic Microbiology 28: 265-278.

Sreekrishna K, Nelles L, Potenz R, Cruse J, Mazzaferro P, Fish W, Fuke M, Holden K, Phelps D, Wood P, and Parker K (1989) High level expression, purification and characterization of recombinant human Tumor Necrosis Factor synthesized in the methylotrophic yeast *Pichia pastoris*. Biochemistry 28: 4117-4125.

Stolze I, Pfannschmidt UB, Freitag P, Wotzlaw C, Rossler J, Frede S, Acker H and Fandrey J (2002) Hypoxia-inducible erythropoietin gene expression in human neuroblastoma cells. Blood 100(7): 2623-2628.

Wen D, Boissel JP, Showers M, Ruch BC and Bunn HF (1994) Erythropoietin structure-function relationships. Identification of functionally

important domains. J Biol Chem 269(26): 22839-22846.

Wojchowski DM, Orkin SH and Sytkowski AJ (1987) Active human erythropoietin expressed in insect cells using a baculovirus vector: a role for N-linked oligosaccharide. Biochim Biophys Acta 910(3): 224-232.

Yadava A and Ockenhouse CF (2003) Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems. Infection and Immunity 71 (9): 4961-4969.

http://www.ncbi.nlm.nih.gov. GenBank database. Accession no. NP_000790. Protein sequence of human erythropoietin

http://www.ncbi.nlm.nih.gov. GenBank database. Accession no. NM_000799. mRNA sequence of human erythropoietin

http://www.kazusa.or.jp/codon/. *Pichis pastoris* codon-usage.

http://www.ncbi.nlm.nih.gov. Online Mendelian Inheritance in Man (OMIM) 133170: Erythropoietin; EPO.