Title: METHODS FOR OPTIMIZING HETEROLOGOUS GENE EXPRESSION

Abstract: The present invention is inter alia concerned with methods for improved recombinant heterologous gene expression. More particularly, the present invention relates e.g. to a method for optimizing a cDNA for recombinant heterologous expression in a cell or to a method for producing a protein by recombinant heterologous expression of a cDNA in a cell. These and other methods of the present invention comprise two essential steps: (i) a step of amending at least one nucleotide of a consensus splice site while maintaining the encoded amino acid sequence and (ii) a step of optimizing codon usage of the modified cDNA in accordance with the codon usage of the cell used for recombinant expression. As a result, a modified cDNA is obtained, which is optimized for the recombinant heterologous expression while still encoding the desired protein in its desired amino acid sequence.


Published:
— with international search report (Art. 21(3))
— with sequence listing part of description (Rule 5.2(a))
Methods for optimizing heterologous gene expression

FIELD OF THE INVENTION

The present invention is in the area of heterologous gene expression. Thus, the present invention relates to a method for obtaining a cDNA for recombinant heterologous expression in a cell, a method for optimizing a cDNA for recombinant heterologous expression in a cell, a method for producing a protein by recombinant heterologous expression of a cDNA in a cell, and a method of optimizing the recombinant heterologous expression of a cDNA in a cell. All of these methods comprise two essential steps, namely a step of amending at least one nucleotide of a consensus splice site in order to arrive at a modified cDNA, and a step of optimizing codon usage of the modified cDNA in accordance with the codon usage of the cell used for recombinant expression in order to arrive at a further modified cDNA. Both steps are carried out such that the encoded amino acid sequence is maintained. The present invention further relates to uses of the modified cDNA gained thereby as well as to uses of the methods of the present invention.

BACKGROUND OF THE INVENTION

Heterologous gene expression is of fundamental importance for basic biology and in medicine for industrial production purposes. In order to optimize transgene
transcription and translation as well as transcript stability, many factors have to be
understood. Engineering of upstream and downstream regulatory sequences,
signal peptides and codon optimization are examples of how to improve protein
production in heterologous production hosts [30].

5

An example of a production platform for heterologous gene expression is the moss
Physcomitrella patens. P. patens has been established as a bioreactor for the
production of complex recombinant biopharmaceuticals and proven its suitability for
large-scale production of recombinant proteins under good manufacturing practice
(GMP) conditions [14,15]. The first moss-produced drug candidate, moss-aGal for
enzyme replacement therapy of Fabry disease [17], has successfully completed Phase
I Clinical Trials [68]. Of course, other production platforms for heterologous gene
expression such as e.g. mammalian cells or insect cells are also well-established.

10

When using the moss P. patens as host cell or production platform, respectively, for
the heterologous production of a protein, such as e.g. a human protein, the native
signal peptide of the protein can be replaced with the moss AP1 signal sequence that
effectively targets the polypeptide chain to the ER, resulting in an efficient secretion
of the recombinant protein [55,56]. In addition, the expression of the protein may e.g.
be driven by the promoter of the PpActin5 gene to obtain high expression [20,57].
These measures are amongst others known to improve the heterologous gene
expression in P. patens and corresponding optimizations are well established for
other production platforms as well.

15

While the above measures result in a higher yield and thus optimize heterologous
gene expression, there is nevertheless a constant need to further improve the
heterologous gene expression in P. patens and other cells serving as production
platforms for heterologous gene expression. It would in particular be desirable to
reduce the bioenergetic costs of recombinant heterologous gene expression by
preventing the formation of non-functional products.
OBJECTS AND SUMMARY OF THE INVENTION

The inventors of the present invention have surprisingly found that splicing takes place if a cDNA comprising only exons is recombinantly expressed in a heterologous cell. In other words, the inventors have found that expression platforms for the production of a protein derived from a different species (i.e. a heterologous expression), wherein the protein is encoded by a cDNA comprising only exons of the different species, nevertheless recognize specific sequences in the cDNA as consensus splice sites such that different protein species are produced. The inventors have furthermore surprisingly found that this splicing (which may also be referred to as “heterosplicing”) can be prevented (i) if the cDNA is modified in that at least one nucleotide of a consensus splice site is amended such that the splice site is destroyed while the encoded amino acid sequence is maintained and (ii) if the codon usage of the further codons is optimized. The use of a cDNA modified in accordance with the above in a heterologous expression system was found by the inventors to result exclusively in the production of the desired full-length protein by the expression platform in a surprisingly high amount.

In the first aspect, the present invention is directed to an in vitro method of obtaining a cDNA for recombinant heterologous expression in a cell, wherein said method comprises the following steps:

a) Providing a cDNA;

b) Identifying nucleotides of consensus splice sites in the cDNA provided in step a);

c) Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;

d) Identifying under-represented codons in cDNA NO:1 obtained in step c);
e) Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2, wherein cDNA NO:2 corresponds to the cDNA for recombinant expression in a cell; wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression; and wherein the cDNA is mammalian cDNA.

In the second aspect, the present invention is directed to an in vitro method of optimizing a cDNA for recombinant heterologous expression in a cell, wherein said method comprises the following steps:

a) Providing a cDNA;
b) Identifying nucleotides of consensus splice sites in the cDNA provided in step a);
c) Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;
d) Identifying under-represented codons in cDNA NO:1 obtained in step c);

e) Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2, wherein cDNA NO:2 corresponds to a cDNA optimized for recombinant expression in a cell; wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression; and wherein the cDNA is mammalian cDNA.

In the third aspect, the present invention is directed to an in vitro method of producing a protein by recombinant heterologous expression of a cDNA in a cell, wherein said method comprises the following steps:
a) Providing a cDNA;
b) Identifying nucleotides of consensus splice sites in the cDNA provided in step a);
c) Amending at least one of the nucleotides identified in step b) such that
   (i) at least one consensus splice site is destroyed while (ii) the
   encoded amino acid sequence is maintained, thereby arriving at
   cDNA NO:1;
d) Identifying under-represented codons in cDNA NO:1 obtained in step c);
e) Replacing the under-represented codons identified in step d) with
   over-represented codons, thereby arriving at cDNA NO:2;
f) Recombinantly expressing cDNA NO:2 obtained in step e) in the cell,
   thereby producing the protein
   wherein the identification of under-represented codons in step d) and the
   replacement with over-represented codons in step e) is based on the codon
   usage of the cell used for recombinant expression; and wherein the cDNA
   is mammalian cDNA.

In one embodiment, step f) is based on a stable integration of cDNA NO:2 into the
genome of said cell. In one alternative embodiment, step f) is based on a transient
transfection of cDNA NO:2 into said cell.

In one embodiment, the produced protein is substantially present inside the cell. In
this case, the protein is preferably an enzyme. The protein or enzyme, respectively,
may be selected from the group consisting of N-acetylglucosamine-1-
phosphotransferase (GNPT), N-acetylglucosamine-1-phosphodiester-alpha-N-
acetylglucosaminidase (NAGPA), mannosyl-phosphate-transferase Mnn6p, UDP-N-
acetylglucosamine-2-epimerase/N-acetylmannosamine kinase (GNE), N-
acetylneuraminic acid phosphate synthase (NANS), N-acetylneuraminic acid
phosphatase (NANP), CMP-N-acetylmuramic acid synthase (CMAS), CMP-sialic
acid transporter (CST), α2,6-sialyltransferase (α2,6-ST), α2,3-sialyltransferase (α2,3-ST), α-N-acetyl-neuraminide alpha-2,8-sialyltransferase 2 (ST8Sia-II), α-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4 (ST8Sia-IV), α1,6-fucosyltransferase (FT), β1,4-galactosyltransferase (GalT4), UDP-GlcNAc 4-epimerase, UDP-GlcNAc/UDP-GalNAc, N-acetylgalactosaminytransferase 2 (GalNAc-T2), serine O-α-galactosyltransferase, Gamma-glutamyl carboxylase (GGCX), Paired basic amino acid residue-cleaving enzyme (PACE), vitamin K epoxide reductase complex subunit 1 (VKORC1), and combinations thereof. A corresponding cell line stably expressing one of the afore-mentioned proteins or enzymes, respectively, or a combination thereof, is in particular useful when producing a target protein, such as e.g. a biopharmaceutical, which is glycosylated or a fully matured target protein, such as e.g. a biopharmaceutical.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the group consisting of N-acetylglucosamine-1-phosphotransferase (GNPT), N-acetylglucosamine-1-phosphodiester-alpha-N-acetylglucosaminidase (NAGPA), mannosyl-phosphate-transferase Mnn6p, UDP-N-acetylglucosamine-2-epimerase/N-acetylmannosamine kinase (GNE), N-acetylenearaminic acid phosphate synthase (NANS), N-acetylenearaminic acid phosphatase (NANP), CMP-N-acetylenearaminic acid synthase (CMAS), CMP-sialic acid transporter (CST), α2,6-sialyltransferase (α2,6-ST), α2,3-sialyltransferase (α2,3-ST), α-N-acetyl-neuraminide alpha-2,8-sialyltransferase 2 (ST8Sia-II), α-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4 (ST8Sia-IV), α1,6-fucosyltransferase (FT), β1,4-galactosyltransferase (GalT4), UDP-GlcNAc 4-epimerase, UDP-GlcNAc/UDP-GalNAc, N-acetylgalactosaminytransferase 2 (GalNAc-T2), serine O-α-galactosyltransferase, and combinations thereof. All of the afore-mentioned enzymes are implicated in producing a glycosylated (target) protein, which is preferably a biopharmaceutical, in particular in plants and more particular in P. patens. A corresponding cell line stably expressing one of the afore-mentioned proteins or enzymes, respectively, or a combination thereof (in particular the combinations set out in the following
embodiments), is in particular useful when producing a target protein, such as e.g. a biopharmaceutical, which is glycosylated.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the group consisting of N-acetylglucosamine-1-phosphotransferase (GNPT), N-acetylglucosamine-1-phosphodiester-alpha-N-acetylglucosaminidase (NAGPA), mannosyl-phosphate-transferase Mnn6p, and combinations thereof. Expressing either GNPT and NAGPA together or relying exclusively on the expression of Mnn6p is a way of producing a mannose-6-phosphate modified glycoprotein, in particular in plants and more particular in *P. patens*.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the group consisting of GalT4, GNE, NANS, NANP, CMAS, CST, α2,6-ST, α2,3-ST, FT, ST8Sia-II, ST8Sia-IV and combinations thereof, wherein it is preferred to select all of the afore-mentioned enzymes together. Expressing GalT4, GNE, NANS, NANP, CMAS, CST, α2,6-ST, α2,3-ST, FT, ST8Sia-II and ST8Sia-IV together is a way of producing a glycoprotein with a humanized *N*-glycosylation pattern, in particular in plants and more particular in *P. patens*.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the group consisting of UDP-GlcNAc 4-epimerase, UDP-GlcNAc/UDP-GalNAc, GalNAc-T2 and combinations thereof, wherein it is preferred to select all of the afore-mentioned enzymes together. Expression of UDP-GlcNAc 4-epimerase, UDP-GlcNAc/UDP-GalNAc and GalNAc-T2 together is a way of producing a glycoprotein with a single GalNAc on serine and/or threonine residues, in particular in plants and more particular in *P. patens*.

In one embodiment thereof, the protein or enzyme, respectively, is serine *O*-α-galactosyltransferase. Expression of serine *O*-α-galactosyltransferase is a way of
producing an \( O \)-glycosylated protein, in particular in plants and more particular in \( P. \) 
\textit{patens}.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the 
group consisting of GGCX, PACE, VKORC1 and combinations thereof. Expression 
of GGCX, PACE and VKORC1 together is a way of producing enzymes that are 
necessary for the maturation of vitamin K-dependent blood coagulation factors 
including Factor II and Factor IX, in particular in plants and more particular in \( P. \) 
\textit{patens}.

In one embodiment, the protein is secreted from the cell into the cell culture medium. 
In this case, the cDNA preferably comprises a sequence coding for a signal peptide 
directing the protein for secretion into the culture medium. Furthermore, in this case, 
it is preferred that the protein is a biopharmaceutical. The protein or 
biopharmaceutical, respectively, may be selected from the group consisting of blood 
coagulation factor II, blood coagulation factor VIII, blood coagulation factor IX, 
moss-aGal, human complement regulator factor H, factor H-related proteins (FHR), 
synthetic/chimeric factor H-related protein, vascular endothelial growth factor 
(VEGF), keratinocyte growth factor (FGF7/KGF), epidermal growth factor (EGF), 
hepatocyte growth factor (HGF), erythropoietin (EPO), asialo-erythropoietin (asialo-
EPO), beta-glucocerebrosidase (GBA), and an immunoglobulin (Ig).

In the \textbf{fourth aspect}, the present invention is directed to an \textit{in vitro} method of 
optimizing the recombinant heterologous expression of a cDNA in a cell, wherein 
said method comprises the following steps:

a) Providing a cDNA;

b) Identifying nucleotides of consensus splice sites in the cDNA 
   provided in step a);

c) Amending at least one of the nucleotides identified in step b) such that 
   (i) at least one consensus splice site is destroyed while (ii) the
encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;
d) Identifying under-represented codons in cDNA NO:1 obtained in step c);
e) Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2;
f) Using cDNA NO:2 obtained in step e) for recombinant expression in a cell, thereby optimizing the recombinant expression of a cDNA in a cell;

wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression; and wherein the cDNA is mammalian cDNA.

In one embodiment, the optimization resides in a reduction of the number of splice variants of the protein. In one embodiment, the optimization resides in an elimination of at least one splice variant of the protein, preferably of all splice variants of the protein.

In one embodiment, the method comprises a step prior to step b), namely the step of recombinantly expressing the cDNA provided in step a) in the cell and determining whether the expression results in splice variants of the encoded protein. The determination whether the expression results in splice variants of the encoded protein may be carried out with standard methods, in particular the methods used in the examples of the present application. Thus, in particular RT-PCR, optionally in combination with sequencing, Western-Blotting, Coomassie-Staining in combination with mass-spectrometry, or the like may be used. RT-PCR is a preferred method.

In the fifth aspect, the present invention is directed to an vitro method of increasing the fraction of the full-length variant of a protein produced by recombinant
heterologous expression of a cDNA in a cell, wherein said method comprises the following steps:

a) Providing a cDNA encoding the full-length variant of a protein;
b) Identifying nucleotides of consensus splice sites in the cDNA provided in step a);
c) Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;
d) Identifying under-represented codons in cDNA NO:1 obtained in step c);
e) Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2;
f) Using cDNA NO:2 obtained in step e) for recombinant expression in a cell, thereby obtaining an increased fraction of the full-length variant of the protein;

wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression; and wherein the cDNA is mammalian cDNA.

In one embodiment, the fraction of the full-length variant of the protein obtained in step f) is increased compared to the fraction of the full-length variant of the protein if the cDNA provided in step a) is used for recombinant expression in a cell.

In one embodiment, the method comprises a step prior to step b), namely the step of recombinantly expressing the cDNA provided in step a) in the cell and determining whether the expression results in splice variants of the encoded protein. The determination whether the expression results in splice variants of the encoded protein may be carried out with standard methods, in particular the methods used in the
examples of the present application. Thus, in particular RT-PCR, optionally in combination with sequencing, Western-Blotting, Coomassie-Staining in combination with mass-spectrometry, or the like may be used. RT-PCR is a preferred method.

5 The following embodiments apply for **all aspects** as outlined above.

In one embodiment, the consensus splice sites are donor and acceptor splice sites.

In one embodiment, the donor and acceptor splice sites follow the GT-AG Rule.

10 In one embodiment, the donor splice site and/or the acceptor splice site is amended in step c). In one embodiment, the donor splice sites and/or the acceptor splice sites are amended in step c). In a preferred embodiment, the donor splice site and the acceptor splice site are amended in step c), and the donor splice sites and the acceptor splice sites are amended in step c), respectively.

15 In one embodiment, the identifying of under-represented codons in cDNA NO:1 of step d) does not include the codons comprising the at least nucleotide amended in step c).

20 In one embodiment, the replacing of step e) does not result in the introduction of consensus splice sites.

In one embodiment, the replacing of step e) results in an increase in the GC-content.

25 In other words, the GC-content of cDNA NO:2 / GC-content of cDNA NO:1 is > 1 in this embodiment.

In one embodiment, the mammalian cDNA is selected from the group consisting of human cDNA, rat cDNA and mouse cDNA. The cDNA may also be a hybrid cDNA comprised of plant cDNA and mammalian cDNA, wherein it is to be understood that
such a hybrid cDNA is still regarded as mammalian cDNA if it comprises mammalian cDNA. It is typical for such a hybrid cDNA that the plant cDNA relates to a specific motif, in particular a signal sequence (ultimately resulting in the efficient secretion of the protein), whereas the mammalian cDNA encodes the protein of interest (that is ultimately secreted). The term “hybrid cDNA” as used herein is synonymous with the term “chimeric cDNA”. The latter term is used more often in recent times to not create confusion with respect to DNA-probes for hybridization.

A hybrid or chimeric cDNA comprising mammalian cDNA (referred to as “mammalian cDNA” herein) is typically a cDNA that is used when producing a protein by recombinant heterologous expression of a cDNA in a cell. Since the cDNA, here the mammalian cDNA, is derived from a different species or cell than the species or cell used for the recombinant gene expression, sequence motifs or elements of the species or cell used for the recombinant gene expression are typically fused to the mammalian cDNA in order to e.g. result in a secretion (in this case, a signal sequence that effectively targets the polypeptide chain to the ER). A result may be a hybrid or chimeric cDNA comprised of mammalian cDNA and plant cDNA. Alternatively, the mammalian cDNA may be fused to sequence motifs or elements not directly derived from the species or cell used for the recombinant gene expression (such as e.g. from *P. patens*) but from a non-mammalian organism, in particular a plant different from the species or cell used for the recombinant gene expression (such as e.g. from *Thuja occidentalis*). Also in this example, the result is a hybrid or chimeric cDNA comprised of mammalian cDNA and plant cDNA.

In one embodiment, the protein is an enzyme or a biopharmaceutical. Said enzyme is preferably selected from the group consisting of N-acetylglucosamine-1-phosphotransferase (GNPT), N-acetylglucosamine-1-phosphodiester-alpha-N-acetylg glucosaminidase (NAGPA), mannosyl-phosphate-transferase Mnn6p, UDP-N-acetylglucosamine-2-epimerase/N-acetylmannosamine kinase (GNE), N-
acetylneuraminic acid phosphate synthase (NANS), N-acetylneuraminic acid phosphatase (NANP), CMP-N-acetylneuraminic acid synthase (CMAS), CMP-sialic acid transporter (CST), α2,6-sialyltransferase (α2,6-ST), α2,3-sialyltransferase (α2,3-ST), α-N-acetyl-neuraminide alpha-2,8-sialyltransferase 2 (ST8Sia-II), α-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4 (ST8Sia-IV), α1,6-fucosyltransferase (FT), β1,4-galactosyltransferase (GalT4), UDP-GlcNAc 4-epimerase, UDP-GlcNAc/UDP-GalNAc, N-acetylgalactosaminyltransferase 2 (GalNAc-T2), serine O-α-galactosyltransferase, Gamma-glutamyl carboxylase (GGCX), Paired basic amino acid residue-cleaving enzyme (PACE), vitamin K epoxide reductase complex subunit 1 (VKORC1), and combinations thereof. A corresponding cell line stably expressing one of the afore-mentioned proteins or enzymes, respectively, or a combination thereof, is in particular useful when producing a target protein, such as e.g. a biopharmaceutical, which is glycosylated or a fully matured target protein, such as e.g. a biopharmaceutical.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the group consisting of N-acetylglucosamine-1-phosphotransferase (GNPT), N-acetylglucosamine-1-phosphodiester-alpha-N-acetylglucosaminidase (NAGPA), mannosyl-phosphate-transferase Mnn6p, UDP-N-acetylglucosamine-2-epimerase/N-acetylmannosamine kinase (GNE), N-acetylneuraminic acid phosphate synthase (NANS), N-acetylneuraminic acid phosphatase (NANP), CMP-N-acetylneuraminic acid synthase (CMAS), CMP-sialic acid transporter (CST), α2,6-sialyltransferase (α2,6-ST), α2,3-sialyltransferase (α2,3-ST), α-N-acetyl-neuraminide alpha-2,8-sialyltransferase 2 (ST8Sia-II), α-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4 (ST8Sia-IV), α1,6-fucosyltransferase (FT), β1,4-galactosyltransferase (GalT4), UDP-GlcNAc 4-epimerase, UDP-GlcNAc/UDP-GalNAc, N-acetylgalactosaminyltransferase 2 (GalNAc-T2), serine O-α-galactosyltransferase, and combinations thereof. All of the afore-mentioned enzymes are implicated in producing a glycosylated (target) protein, which is preferably a biopharmaceutical, in particular in plants and more particular in *P. patens*. A corresponding cell line stably
expressing one of the afore-mentioned proteins or enzymes, respectively, or a combination thereof (in particular the combinations set out in the following embodiments), is in particular useful when producing a target protein, such as e.g. a biopharmaceutical, which is glycosylated.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the group consisting of N-acetylglucosamine-1-phosphotransferase (GNPT), N-acetylglucosamine-1-phosphodiester-alpha-N-acetylglucosaminidase (NAGPA), mannosyl-phosphate-transferase Mnn6p, and combinations thereof. Expressing either GNPT and NAGPA together or relying exclusively on the expression of Mnn6p is a way of producing a mannose-6-phosphate modified glycoprotein, in particular in plants and more particular in *P. patens*.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the group consisting of GalT4, GNE, NANS, NANP, CMAS, CST, \( \alpha_2,6-ST \), \( \alpha_2,3-ST \), FT, ST8Sia-II, ST8Sia-IV and combinations thereof, wherein it is preferred to select all of the afore-mentioned enzymes together. Expressing GalT4, GNE, NANS, NANP, CMAS, CST, \( \alpha_2,6-ST \), \( \alpha_2,3-ST \), FT, ST8Sia-II and ST8Sia-IV together is a way of producing a glycoprotein with a humanized \( N \)-glycosylation pattern, in particular in plants and more particular in *P. patens*.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the group consisting of UDP-GlcNAc 4-epimerase, UDP-GlcNAc/UDP-GalNAc, GalNAc-T2 and combinations thereof, wherein it is preferred to select all of the afore-mentioned enzymes together. Expression of UDP-GlcNAc 4-epimerase, UDP-GlcNAc/UDP-GalNAc and GalNAc-T2 together is a way of producing a glycoprotein with a single GalNAc on serine and/or threonine residues, in particular in plants and more particular in *P. patens*. 
In one embodiment thereof, the protein or enzyme, respectively, is serine O-α-galactosyltransferase. Expression of serine O-α-galactosyltransferase is a way of producing an O-glycosylated protein, in particular in plants and more particular in *P. patens*.

In one embodiment thereof, the protein or enzyme, respectively, is selected from the group consisting of GGCX, PACE, VKORC1 and combinations thereof. Expression of GGCX, PACE and VKORC1 together is a way of producing enzymes that are necessary for the maturation of vitamin K-dependent blood coagulation factors including Factor II and Factor IX, in particular in plants and more particular in *P. patens*.

Said biopharmaceutical is preferably selected from the group consisting of blood coagulation factor II, blood coagulation factor VIII, blood coagulation factor IX, moss-aGal, human complement regulator factor H, factor H-related proteins (FHR), synthetic/chimeric factor H-related protein, vascular endothelial growth factor (VEGF), keratinocyte growth factor (FGF7/KGF), epidermal growth factor (EGF), hepatocyte growth factor (HGF), erythropoietin (EPO), asialo-erythropoietin (asialo-EPO), beta-glucocerebrosidase (GBA), and an immunoglobulin (Ig).

The identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell (or the species of the cell, respectively) used for recombinant expression. In other words, the under-represented and over-represented codons are derived from the codon usage in the cell (or the species of the cell, respectively) used for recombinant expression.

In a particular preferred embodiment, the cell is a plant cell. Said plant cell may be an isolated cell, a singularized cell, a cell in or of a plant tissue, preferably a tissue selected from the group consisting of callus, protonema, phloem, xylem, mesophyll, trunk, leaves, thallus, chloronema, caulonema, rhizoid or gametophore and a cell in a
plant organism. Said plant cell is preferably a cell from a plant selected from the group consisting of bryophyta, liverwort, hornwort and an aquatic plant or alga.

Said bryophyta is preferably moss, more preferably a moss of the class bryopsida. It can be particularly preferred in the present invention that said bryopside is *Physcomitrella patens* such that said cell is a *P. patens* cell.

Said aquatic plant is preferably an aquatic plant from the genus selected from the group consisting of *Lemna*, *Spirodea*, *Landoltia*, *Wolffia* and *Wolfiella*.

In one embodiment, the cell is a mammalian cell. Preferably, said mammalian cell is selected from the group consisting of a CHO cell, a HeLa cell, a HEK cell, a U2OS cell, an A549 cell, a HT1080 cell, a CAD cell, a P19 cell, a NIH 3T3 cell, a L929 cell, a N2a cell, a MCF-7 cell, a Y79 cell, a SO-Rb50 cell, a Hep G2 cell, a DUKX-X11 cell, a J558L cell, and a BHK cell.

In one embodiment, the cell is an insect cell. Preferably, said insect cell is selected from the group consisting of a Sf9 cell, a Mimic Sf9 cell, a Sf21 cell, a High Five cell, a S2 cell, and a D.Mel2 cell.

In one embodiment, the cDNA is a mammalian cDNA, preferably a human cDNA; and the cell is a plant cell, preferably a bryopside cell, more preferably a *P. patens* cell. In one embodiment, the cDNA is a mammalian cDNA, preferably a human cDNA; the cell is a plant cell, preferably a bryopside cell, more preferably a *P. patens* cell; and the donor and acceptor splice sites follow the GT-AG Rule. In one embodiment, the cDNA is a mammalian cDNA, preferably a human cDNA; the cell is a plant cell, preferably a bryopside cell, more preferably a *P. patens* cell; the donor and acceptor splice sites follow the GT-AG Rule; and the under-represented and over-represented codons are derived from the codon usage in said plant cell, preferably in said bryopside cell, more preferably in said *P. patens* cell. In the afore-
mentioned embodiments, it can be particularly preferred that said cDNA encodes FIX.

In the sixth aspect, the present invention is directed to the use of a cDNA obtained according to the method of the first aspect for recombinant heterologous expression in a cell.

In the seventh aspect, the present invention is directed to the use of a cDNA optimized according to the method of the second aspect for recombinant heterologous expression in a cell.

In the eighth aspect, the present invention is directed to the use of the method of the third aspect in the generation of a host cell for heterologous recombinant expression of a mammalian cDNA, wherein said host cell is capable of carrying out at least one posttranslational modification found in the mammal. This in particular applies to a method of the third aspect, where a protein or a combination of proteins, respectively, is produced inside the cell that is/are (an) enzyme(s) responsible for the posttranslational modification. The posttranslational modification may e.g. be glycosylation, sialylation or the like.

In the ninth aspect, the present invention is concerned with a protein produced by the method according to the third aspect.

The embodiments of the first five aspects apply to the sixth, seventh, eighth and ninth aspect as well, in particular when it comes to the cDNA and the cell used for recombinant heterologous expression.
DESCRIPTION OF THE FIGURES

Figure 1 – Expression of FIX in *Physcomitrella patens*.
A: Total RNA from protonema cultures of 4 FIX-transgenic lines (FGP) and the
parental line Δxt/ft was prepared and used for RT-PCR using the primers FIXfwdB
and FIXrevB.
B: RT-PCR of FIX using the primers FIXfwdB and FIXrevB from gametophores of
the same lines.
C: RT-PCR of FIX with mRNA extracted from transiently transfected (FIX) as well
as non-transfected (Δxt/ft) cells.
M: 1 kb Marker (Thermo Fisher Scientific), -: without reverse transcription, +: with
reverse transcription, CO: Water control.

Figure 2 – Schematic representation of FIX transcripts and multiple sequence
alignment of predicted FIX protein isoforms based on transcripts detected in
*Physcomitrella patens*.
A: Full-length FIX transcript (A.I), FIX variant with the deletion of a 44 aa long
domain in the CDS for the heavy chain (A.II), FIX variant with deletions in the CDS
for both light and heavy chain (A.III), FIX variant with deletions in the CDS for light
chain, activation peptide and heavy chain (A.IV), FIX variant missing the complete
CDS for the activation peptide, almost the complete light chain and a 41 aa long
domain in the heavy chain (A.V).
“*” indicates the usage of a slightly earlier donor site, which is -9 bp upstream of the
site for variants II, III and IV; S: PpAPI signal peptide, P: propeptide, AP: activation
peptide, H: 8x His tag; “1” to “6” indicates splice sites, see Figure 8.
B: Alignment was performed and formatted for publication with MUSCLE via the
"MSA" package for R [49,50]. The protein sequences have the following SEQ IDs:
Complete FIX: SEQ ID NO:39; FIX Variant 1: SEQ ID NO:40; FIX Variant 2: SEQ
ID NO:41; FIX Variant 3: SEQ ID NO:42; FIX Variant 4: SEQ ID NO:43.
Figure 3 – Schematic representation of donor and acceptor sites within the FIX sequence and consensus sequence motifs of donor and acceptor splice sites recovered from all 87,533 *P. patens* transcripts.

A: Donor and acceptor sites. Four different splicing motifs were identified after sequencing FIX variants. Left side shows donor site and right side shows acceptor site. "-1" represents last nucleotide of exon in donor site (left side) or last nucleotide of intron in acceptor site (right side). "1" to "6" indicates splice sites, see Figure 8. The SEQ IDs of the sequences depicted in this Figure are as follows: splice site 4 is SEQ ID NO: 4; splice site 1 is SEQ ID NO: 1; splice site 6 is SEQ ID NO: 6; splice site 2 is SEQ ID NO: 2; splice site 3 is SEQ ID NO: 3; splice site 5 is SEQ ID NO: 5.

B, C: Consensus genomic sequence motifs of donor (B) and acceptor splice sites (C) extracted from all 32,926 *P. patens* protein-coding genes. Probability and height of individual letters correspond to base frequencies at each position. The SEQ IDs of the sequences depicted in Figures B and C are as follows: B: SEQ ID NO: 7; C: SEQ ID NO: 8.

Figure 4 – Analysis of moss-produced FIX protein variants.

A: Immunodetection of extracellular FIX protein produced by two transfections (anti-FIX Ab 1:5,000). Culture supernatant of non-transfected cells was used as negative control (Neg). Exposure time: 10 sec. T1, T2: Two different transient transfections with the FIX cDNA-based expression plasmid.

B: Full-length mossFIX amino acid sequence. Amino acids marked in yellow were identified in the MS analysis. Blue boxes indicate the two parts of the peptide VVGGEDAKPGEHNIETETEQKR (SEQ ID NO:11) which was also identified in MS analysis and represents a heterospecific product. The SEQ IDs of the sequences depicted in this Figure are as follows: complete protein sequence of FIX protein as indicated: SEQ ID NO:12; 1<sup>st</sup> yellow box: SEQ ID NO:13; 2<sup>nd</sup> yellow box: SEQ ID NO:14; 3<sup>rd</sup> yellow box: SEQ ID NO:15; 4<sup>th</sup> yellow box: SEQ ID NO:16; 1<sup>st</sup> blue box: SEQ ID NO: 9; 2<sup>nd</sup> blue box: SEQ ID NO:10; 5<sup>th</sup> yellow box: SEQ ID NO:17; 6<sup>th</sup>
yellow box: SEQ ID NO:18; 7th yellow box: SEQ ID NO:19; 8th yellow box: SEQ ID NO:20.

Figure 5 – Prevention of Heterospling of the human FIX CDS in

Physcomitrella patens.

A: RT-PCR of FIX RNA from cells that were transiently transfected with aspFIX and non-transfected cells (WT). M: 1 kb Marker (Thermo Fisher Scientific), -RT: without reverse transcription, +RT: with reverse transcription, P: PCR with aspFIX plasmid used as positive control.

B: Splicing motifs identified after sequencing the FIX variant “aspFIX”. The left side shows the donor site and the right side the acceptor site. “-1” represents the last nucleotide of the exon at the donor site or the intron at the acceptor site, respectively. The donor site, site 7, is upstream of donor site 1. The acceptor site corresponds to acceptor site 5 except for the exchange of A to G in order to arrive at aspFIX and is therefore referred to as acceptor site 5’. “5’” and “7” indicates splice sites, see Figure 8. The SEQ IDs of the sequences depicted in this Figure are as follows: Splice site 7 is SEQ ID NO:27; splice site 5’ is SEQ ID NO: 28;

C: Schematic representation of complete FIX (upper) and the additional shorter variant (lower). S: signal peptide, P: propeptide, AP: activation peptide, H: 8x His tag. “5’” and “7” indicates splice sites, see Figure 8.

D: RT-PCR of FIX RNA from cells that were transiently transfected with optiFIX and non-transfected cells (WT), respectively. M: 1 kb Marker (Thermo Fisher Scientific), -RT: without reverse transcription, RT: with reverse transcription.

E: Immunodetection of extracellular FIX protein produced in two transfections (T1, T2), with anti-FIX Ab (1:5,000). Culture supernatant of non-transfected cells was used as negative control (Neg). Exposure time: 1 sec. T1, T2: Two different transient transfections with the optiFIX plasmid.
Figure 6 – 3D rendered confocal Z-stacks of moss protoplasts transfected with FIX-Citrine (A) and optiFIX-Citrine (B) and comparison of normalized mean voxel intensity in 3 images recorded for each construct (C).

A: Transient overexpression of FIX-Citrine yields fluorescence distributed in a sparse network.
B: Transient overexpression of optiFIX-Citrine displays fluorescence signal in a much greater volume of the cell.
C: After normalization the mean signal intensity of optiFIX-Citrine is about 9 times that of FIX-Citrine.

Figure 7 – Multiple sequence alignment of FIX, aspFIX and optiFIX CDS.
The encoded amino acid sequence of FIX (FIX aa) was added for better visualization. Alignment was performed and formatted for publication with MUSCLE via the "MSA" package for R [49,50]. The sequences have the following SEQ IDs: FIX aa: SEQ ID NO:39; FIX: SEQ ID NO:21; aspFIX: SEQ ID NO:22; optiFIX: SEQ ID NO:23.

Figure 8 – Multiple sequence alignment of FIX, aspFIX and optiFIX CDS with splice sites indicated in boxes.
Based on Figure 7, the splice sites as indicated in Figures 2A and 5C with the sequences as given in Figure 3A and 5B are shown in boxes in the alignment. Splice sites 1, 4 and 7 are donor splice sites, whereas splice sites 2, 3, 5, 5' and 6 are acceptor splice sites. It is noted that boxes 5 and 6 as well as 1 and 7 overlap such that boxes 6 and 7 are depicted above the aligned sequences. The sequences have the following SEQ IDs: FIX aa: SEQ ID NO:39; FIX: SEQ ID NO:21; aspFIX: SEQ ID NO:22; optiFIX: SEQ ID NO:23.

Figure 9 – Sequence alignment of CST and optiCST CDS.
Alignment was performed and formatted for publication with MUSCLE via the "MSA" package for R [49,50]. The sequences have the following SEQ IDs: CST: SEQ ID NO:44; optiCST: SEQ ID NO:45.

5 **Figure 10 – Fluorescence microscopy of moss protoplasts transfected with CST-Citrine (A) and optiCST-Citrine.**
A: Transient overexpression of CST-Citrine yields a fluorescence signal. The intensity of this signal is lower compared to the signal in B.
B: Transient overexpression of optiCST-Citrine displays a higher fluorescence signal compared to A.
The scale bar in A and B is 20 μm.

10 **Figure 11 – Sequence alignment of FH and optiFH CDS.**
Alignment was performed and formatted for publication with MUSCLE via the "MSA" package for R [49,50]. The sequences have the following SEQ IDs: FH: SEQ ID NO:46; optiFH: SEQ ID NO:47. It is noted that both sequences (also in SEQ ID NO:46 and SEQ ID NO:47) contain a 5’ sequence encoding a signal peptide (the first 78 bases). In the FH-sequence, this sequence encodes the native signal peptide of the H1-gene from *Thuja occidentalis*, whereas in the optiFH-sequence, this native H1-encoding sequence was optimized according to the present invention, therefore resulting in a slightly different sequence compared to the native H1-encoding sequence.

20 **Figure 12 – 3D rendered confocal Z-stacks of moss protoplasts transfected with FH-Citrine (A) and optiFH-Citrine.**
A: Transient overexpression of FH-Citrine does not results in a detectable fluorescence signal.
B: Transient overexpression of optiFH-Citrine displays a high fluorescence signal.
The scale bar in A and B is 3 μm.
DETAILED DESCRIPTION OF THE INVENTION

Before the present invention is described in more detail in the example section, the underlying findings are described and definitions are introduced.

1. Underlying findings and definitions

The present invention is based on the surprising finding that heterosplicing, i.e. splicing in a cell used for recombinant heterologous expression of a cDNA, can successfully be prevented if (i) at least one consensus splice site is destroyed by amending at least one nucleotide of said site while maintaining the encoded amino acid sequence and (ii) optimizing the codon usage in accordance with the cell used for recombinant heterologous expression of the cDNA, wherein the two steps result in a modified cDNA, which is optimized for the heterologous expression in said cell.

As used in the specification and the claims, the singular forms of “a” and “an” also include the corresponding plurals unless the context clearly dictates otherwise.

The term “about” in the context of the present invention denotes an interval of accuracy that a person skilled in the art will understand to still ensure the technical effect of the feature in question. The term typically indicates a deviation from the indicated numerical value of ±10% and preferably ±5%.

It needs to be understood that the term “comprising” is not limiting. For the purposes of the present invention, the term “consisting of” is considered to be a preferred embodiment of the term “comprising”. If hereinafter a group is defined to comprise at least a certain number of embodiments, this is also meant to encompass a group which preferably consists of these embodiments only.
The term "DNA" is the usual abbreviation for deoxyribonucleic acid. It is a nucleic acid molecule, i.e. a polymer consisting of nucleotide monomers. These nucleotides are usually deoxy-adenosine-monophosphate, deoxy-thymidine-monophosphate, deoxy-guanosine-monophosphate and deoxy-cytidine-monophosphate monomers or analogs thereof which are – by themselves – composed of a sugar moiety (deoxyribose), a base moiety and a phosphate moiety, and polymerize by a characteristic backbone structure. The backbone structure is, typically, formed by phosphodiester bonds between the sugar moiety of the nucleotide, i.e. deoxyribose, of a first and a phosphate moiety of a second, adjacent monomer. The specific order of the monomers, i.e. the order of the bases linked to the sugar/phosphate-backbone, is called the DNA-sequence. DNA may be single stranded or double stranded. In the double stranded form, the nucleotides of the first strand typically hybridize with the nucleotides of the second strand, e.g. by A/T-base-pairing and G/C-base-pairing.

The term "RNA" is the usual abbreviation for ribonucleic acid. It is a nucleic acid molecule, i.e. a polymer consisting of nucleotide monomers. These nucleotides are usually adenosine-monophosphate, uridine-monophosphate, guanosine-monophosphate and cytidine-monophosphate monomers or analogs thereof, which are connected to each other along a so-called backbone. The backbone is formed by phosphodiester bonds between the sugar, i.e. ribose, of a first and a phosphate moiety of a second, adjacent monomer. The specific order of the monomers, i.e. the order of the bases linked to the sugar/phosphate-backbone, is called the RNA-sequence.

The term "messenger RNA" or "mRNA" refers to one type of RNA molecule. In vivo, transcription of DNA usually results in the so-called premature RNA which has to be processed into so-called messenger RNA, usually abbreviated as mRNA. Processing of the premature RNA, e.g. in eukaryotic organisms, comprises a variety of different posttranscriptional modifications such as in particular splicing, but also 5'- capping, polyadenylation, export from the nucleus or the mitochondria and the like. The sum of these processes is also called maturation of mRNA. The mature
messenger RNA provides the nucleotide sequence that may be translated into an amino acid sequence of a particular peptide or protein via the genetic code, i.e. the corresponding triplets or codons. Typically, a mature mRNA comprises a 5' cap, a 5'UTR, an open reading frame, a 3'UTR and a poly(A) or a poly(C) sequence.

The term “cDNA” as used herein means a coding sequence of a gene, where no introns of the cell, the cDNA is derived from, are present. In other words, all introns have been removed and the sequence is exclusively comprised of exons when looked at from the perspective of the cell, the cDNA is derived from. “cDNA” as used herein may alternatively be referred to as “cDNA encoding a protein” or “cDNA coding for a protein”. The encoded protein may be a full-length or truncated version of a naturally occurring protein or it may be a single domain of a naturally occurring protein. Alternatively, the encoded protein may comprise several domains of a naturally occurring protein but may not be identical to the full length protein, e.g. it may comprise less domains than the full-length protein. The encoded protein may also comprise different domains of at least two naturally occurring proteins (which are preferably from the same species, e.g. from a mammal, in particular from a human being). The at least two naturally occurring proteins may be from unrelated proteins or from the same protein family, i.e. the at least two naturally occurring proteins are in this case related, such as e.g. isoforms of a specific protein. If domains of different protein are fused, one may also refer to a “synthetic” fusion protein. An example here is the synthetic fusion protein MFHR1 that comprises the regulatory domains of the human complement factor H (FH) and the human C5 convertase/C5b-9 inhibitory fragment of the FH-related protein 1, see [69] and [70].

The cDNA encoding the afore-mentioned exemplary synthetic fusion protein is still to be regarded as mammalian cDNA, more specifically human cDNA, since the encoded domains are derived from mammalian, more specifically human, proteins, as will also be explained in the next paragraph.
The term “mammalian cDNA” as used herein not only applies for mammalian cDNA that consists entirely of mammalian cDNA but also to cDNA that comprises mammalian cDNA (i.e. mammalian cDNA encoding a protein or at least one domain thereof). The cDNA may, in addition to the mammalian cDNA, comprise further cDNA, such as in particular DNA encoding peptide-linkers if several domains are fused and/or signal peptides. A particularly preferred signal peptide encoding DNA is cDNA encoding a signal sequence that effectively targets the encoded mammalian protein for secretion, wherein such a signal sequence is typically cleaved off from the encoded mammalian protein after it has successfully reached its destination inside the cell. This is usually the Endoplasmatic Reticulum, where the signal peptide is cleaved off. Examples of corresponding mammalian cDNAs are the human Factor IX cDNA and the human Factor H cDNA of the examples of the present application that are both synthesized as N-terminal translational fusions to signal peptides, wherein the signal peptides are two different signal peptides from plants. In the case of FIX, the signal-peptide encoding cDNA encodes the PpAP1 signal peptide from P. patens. In the case of FH, the signal-peptide encoding cDNA encodes the signal peptide of the H1 gene from T. occidentalis. As the products to be produced are clearly the human (i.e. mammalian) proteins Factor IX and Factor H, the cDNAs are understood to be mammalian cDNAs although both contain a small plant sequence, namely the signal peptide coding sequence of PpAP1 and the signal peptide coding sequence of H1, respectively.

The term “recombinant expression” as used herein means that a protein is produced by a cell or by an organism comprising at least one cell. In order to do so, the gene expression system of the cell or organism is used and typically manipulated such that large amounts of a recombinant gene (i.e. the cDNA) are expressed. The cDNA, which may be integrated into the genome of the cell, is transcribed into mRNA, followed by the translation of mRNA into a polypeptide, which is folded into the protein and may further be targeted to subcellular or extracellular locations.
The term “consensus splice sites” as used herein refers to i) at least one donor splice site and (ii) at least one acceptor splice site in the cDNA recognized by the cell used for recombinant gene expression. It is important to understand that this is not in contradiction to the definition of the term “cDNA” as defined above, as will be explained in the following. Thus, since the present invention deals with heterologous gene expression, the cDNA is derived from a different species or cell than the species or cell used for the recombinant gene expression. This means that the cDNA does not comprise any introns in terms of the cell or species, it is derived from. If e.g. a human cDNA is used, it does not comprise any human introns, such as in particular introns known from the upstream primary human RNA transcript, which is still subject to splicing in a human cell or species. However, since the human cDNA is expressed in a different cell or organism, e.g. the moss *P. patens*, this different cell or organism unexpectedly treats specific sequence motifs in the human cDNA without human introns as consensus splice sites (in the present example as *P. patens* splice sites) with the result that splicing takes place in the production cell or organism such that one or more parts of the sequence are removed as alleged introns. This may have a significant negative effect on the heterologous expression in that not only the desired full-length but also several other versions of the protein are produced. The inventors refer to this effect as “heterosplicing”.

It is preferred that the “consensus splice sites” follow the GT-AG Rule, meaning that the sequence treated as intron starts with GT (on a DNA-level, i.e. GU on an RNA-level) and ends with AG. The donor splice site comprises the GT (or GU on an RNA-level), whereas the acceptor splice site comprises the AG. The sequences of the consensus splice sites may thus be identified by analyzing the cDNA sequence for sequence regions starting with GT (on a DNA-level, i.e. GU on an RNA-level) and ending with AG.

If the cell is the moss *P. patens*, the sequences of the consensus splice sites are depicted in Figure 3B (donor splice site with the consensus sequence shown in SEQ
ID NO:7) and Figure 3C (acceptor splice site with the consensus sequence shown in SEQ ID NO:8). Such consensus splice sites can easily be identified manually.

On a general level, the process of splicing as discussed herein (which may also be referred to as “alternative splicing” or “AS”) is a fundamental regulatory process that contributes to proteome expansion [32]. AS can produce multiple mRNAs from the same gene through the preference of splice sites during pre-mRNA splicing. It is modulated based on sequence motifs in the pre-mRNA, the interactions between RNA-binding proteins and splice sites, different cell types and environmental signals [33,34]. Splicing is carried out by a large ribonucleoprotein (RNP) complex, the spliceosome, which brings selected exons together by two transesterification reactions [35]. Over the course of splicing, uridine (U)-rich small nuclear RNPs (snRNPs) together with non-snRNP splicing factors, and serine/arginine-rich (SR) proteins participate in the recognition of 5' and 3'-splice sites as well as the branch site [36]. In addition to its role in proteome expansion, AS controls transcript levels by generating unstable mRNA isoforms that may activate nonsense-mediated mRNA decay (NMD) [37–39].

The term “amending a nucleotide” means that a specific nucleotide of the cDNA sequence, i.e. either an A, T, G or C is changed into a different nucleotide selected from A, T, G and C. Thus, an A may e.g. be amended into a T or a G.

As is derivable from the aspects of the present invention, the amendment must be made such that the following two results are achieved: (i) the consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained.

As regards (i) and given the above on preferred splice sites following the GT-AG Rule, it is preferred that either the G and/or the T of the donor splice site is amended. The decision on which nucleotide(s) will be amended is dependent on its/their position(s) in the codon as there should be no change in the amino acid sequence (see
(ii) below). Amending the G and/or T of the donor splice site (optionally in combination with further amendments of nucleotides in the consensus site) destroys the donor splice site such that it is no longer recognized by the cell or organism as donor splice site. Furthermore, it is preferred that either the A and/or the G of the acceptor splice site is amended. The decision on which nucleotide(s) will be amended is dependent on its/their position(s) in the codon as there should be no change in the amino acid sequence (see (ii) below). Amending the A and/or G of the acceptor splice site (optionally in combination with further amendments of nucleotides in the consensus site) destroys the acceptor splice site such that it is no longer recognized by the cell or organism as acceptor splice site. As noted above, it is most preferred to not only amend at least one nucleotide of the donor splice site such that the donor splice site is destroyed but to also amend at least one nucleotide of the acceptor splice site such that the acceptor splice site is destroyed as well – in this case, all consensus splice sites would be destroyed.

As regards (ii), the amendment of at least the G and/or T of the donor splice site and/or of at least the A and/or G of the acceptor splice site must be carried out such that the encoded amino acid sequence is maintained. This means that the amendment must be carried out such that the genetic code is not altered in terms of the specific amino acid encoded by a codon (i.e. a nucleotide triplet) comprising the nucleotide that is intended to be amended. The skilled person understands that the frame of the cDNA is to be taken into account when analyzing the position of the respective nucleotide that is intended to be amended in a codon.

The term "codon usage" as used herein can be defined as the nonrandom and selective usage of synonymous codons due to the fact that eighteen amino acids are encoded by more than one codon. The selection of one over another for the same amino acid differs not only between species but can even differ within different tissues [43]. Thus, expressing genes in a heterologous species without host-specific codon optimization might lead to suboptimal ribosome use and eventually reduced
protein biosynthesis rates. Importantly, as can be derived from the example section of the present application, the use of under-represented codons even has an impact on splicing, namely that alternative sequences may be used as splice sites despite the destruction of at least one of the consensus splice sites identified in step b). In this respect, it is noted that the codon optimization in the present case seems not only have an impact on ribosome use but, even more importantly, seems to eliminate cryptic introns, most likely due to an increase in the overall GC-content.

Before this background, the term “under-represented codons” as used herein means that codons are used in the production cell or species, for which a lower number of tRNAs (or even no tRNA at all) is present in the cell or species. This results inter alia in the suboptimal ribosome use as mentioned above. To the contrary, “over-represented codons” as used herein means that codons are used in the production cell or species, which are typically used by the cell or species such that a high number of corresponding tRNAs is present in the cell or species. Typical under-represented and over-represented codons for eight amino acids are e.g. shown in different expression cells or production systems in Table 1 in the example section below. The codon usage in a specific cell or organism to be used for recombinant heterologous expression can generally easily be determined by a skilled person, namely e.g. by Jcat (http://www.jcat.de), Optimizer (http://genomes.urv.es/OPTIMIZER/), and Codon Optimization On-Line (http://cool.syncti.org/index.php).

In order to emphasize that the step of replacing under-represented codons with over-represented codons does not result in an amended encoded amino acid sequence, step e) of the methods of the present invention may alternatively be referred to as replacing the under-represented codons identified in step d) with over-represented codons coding for the same amino acids, thereby arriving at cDNA NO:2.

The term “biopharmaceutical” as used herein refers to a protein that is used in the medical field. An example for such a biopharmaceutical is e.g. human blood-clotting
factor IX (FIX), which is used for the treatment of Hemophilia B. Hemophilia B is an inherited bleeding disorder caused by the absence or abnormal levels of functional FIX, and recombinant FIX is used to replace the absent or non-functional FIX.

Further preferred embodiments of the present application relate to:

1. An in vitro method of obtaining a cDNA for recombinant heterologous expression in a cell, wherein said method comprises the following steps:
   a) Providing a cDNA;
   b) Identifying nucleotides of consensus splice sites in the cDNA provided in step a);
   c) Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;
   d) Identifying under-represented codons in cDNA NO:1 obtained in step c);
   e) Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2, wherein cDNA NO:2 corresponds to the cDNA for recombinant expression in a cell.

2. An in vitro method of optimizing a cDNA for recombinant heterologous expression in a cell, wherein said method comprises the following steps:
   a) Providing a cDNA;
   b) Identifying nucleotides of consensus splice sites in the cDNA provided in step a);
   c) Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the
encoded amino acid sequence is maintained, thereby arriving at
cDNA NO:1;
d) Identifying under-represented codons in cDNA NO:1 obtained in step
c);
e) Replacing the under-represented codons identified in step d) with
over-represented codons, thereby arriving at cDNA NO:2,
wherein cDNA NO:2 corresponds to a cDNA optimized for recombinant
expression in a cell.

3. An in vitro method of producing a protein by recombinant heterologous
expression of a cDNA in a cell, wherein said method comprises the following
steps:
   a) Providing a cDNA;
   b) Identifying nucleotides of consensus splice sites in the cDNA
      provided in step a);
   c) Amending at least one of the nucleotides identified in step b) such that
      (i) at least one consensus splice site is destroyed while (ii) the
      encoded amino acid sequence is maintained, thereby arriving at
      cDNA NO:1;
   d) Identifying under-represented codons in cDNA NO:1 obtained in step
c);
   e) Replacing the under-represented codons identified in step d) with
      over-represented codons, thereby arriving at cDNA NO:2;
   f) Recombinantly expressing cDNA NO:2 obtained in step e) in the cell,
      thereby producing the protein.

4. An in vitro method of optimizing the recombinant heterologous expression of
a cDNA in a cell, wherein said method comprises the following steps:
   a) Providing a cDNA;
b) Identifying nucleotides of consensus splice sites in the cDNA provided in step a);

c) Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;

d) Identifying under-represented codons in cDNA NO:1 obtained in step c);

e) Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2;

f) Using cDNA NO:2 obtained in step e) for recombinant expression in a cell, thereby optimizing the recombinant expression of a cDNA in a cell.

5. An *in vitro* method of increasing the fraction of the full-length variant of a protein produced by recombinant heterologous expression of a cDNA in a cell, wherein said method comprises the following steps:

a) Providing a cDNA encoding the full-length variant of a protein;

b) Identifying nucleotides of consensus splice sites in the cDNA provided in step a);

c) Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;

d) Identifying under-represented codons in cDNA NO:1 obtained in step c);

e) Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2;
f) Using cDNA NO:2 obtained in step e) for recombinant expression in a cell, thereby obtaining an increased fraction of the full-length variant of the protein.

6. The method according to any one of embodiments 3 to 5, wherein said method comprises a step prior to step b), namely the step of recombinantly expressing the cDNA provided in step a) in the cell and determining whether the expression results in splice variants of the encoded protein.

7. The method according to any one of the preceding embodiments, wherein the consensus splice sites are donor and acceptor splice sites.

8. The method according to embodiment 7, wherein the donor and acceptor splice sites follow the GT-AG Rule.

9. The method according to any one of the preceding embodiments, wherein the cDNA is mammalian cDNA, preferably mammalian cDNA selected from the group consisting of human cDNA, rat cDNA and mouse cDNA.

10. The method according to any one of the preceding embodiments, wherein said cell is selected from the group consisting of a plant cell, a mammalian cell and an insect cell.

11. The method according to any one of the preceding embodiments, wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression.

12. The method according to any one of the preceding embodiments, wherein said cDNA is a mammalian cDNA, preferably a human cDNA; wherein said cell is a plant cell, preferably a *P. patens* cell; wherein the donor and acceptor splice sites follow the GT-AG Rule; and wherein the under-represented and
over-represented codons are derived from the codon usage in said plant cell, preferably in the *P. patens* cell.

13. Use of a cDNA obtained according to the method of embodiment 1 for recombinant heterologous expression in a cell.

14. Use of a cDNA optimized according to the method of embodiment 2 for recombinant heterologous expression in a cell.

15. Use of the method according to embodiment 3 in the generation of a host cell for recombinant heterologous expression of a mammalian cDNA, wherein said host cell is capable of carrying out at least one posttranslational modification found in the mammal.

2. **Examples**

**Example 1:** Expression of the human blood clotting factor IX gene in *P. patens* results in the expected full length product as well as four discrete shorter versions thereof.

The aim of this example was to produce the human protein “blood clotting factor IX” (referred to in the following as “FIX”) in the moss *P. patens*. In order to achieve this, a human FIX-encoding, cDNA-based expression construct comprising the complete FIX coding sequence (CDS) without any introns (see example 7 for details) was transfected into the parental moss line Δxt/ft [20]. 49 transgenic human FIX-expressing lines were generated.

In order to validate the completeness of the FIX transcripts in the 49 transgenic moss lines, reverse transcription PCR (RT-PCR) using total cellular RNA with the primers FIXfwdB and FIXrevB (see example 7 for details) was carried out. In addition to the
full-length transcript, several smaller products were detected in cDNA resulting from RT-PCR from the transgenic moss lines (Figure 1). Discrete PCR products of different sizes were detected in the juvenile protonema tissue of transgenic lines, ranging from about 700 to 1370 bp (Figure 1A). Sequencing of the PCR products revealed that five of the observed bands were FIX-specific products with a length of 1371 bp, 1240 bp, 895 bp, 769 bp, and 691 bp, respectively. The FIX-product with a length of 1371 bp corresponds to the full length FIX-product, whereas the other four products are shorter versions thereof. The sequences of the five products were identical in all 49 transgenic moss lines.

To analyze this phenomenon across different tissues, the above procedure was repeated with RNA isolated from adult plants (gametophores) derived from the 49 transgenic human FIX-expressing lines. FIX transcripts detected in gametophores were consistent with those from protonema cultures (Figure 1B). In addition to direct sequencing, the RT-PCR products were cloned into the pJET1.2blunt vector. This was done as a double check for sequences of the PCR products. Colonies with different insert sizes were selected and sequenced. The initial results, i.e. the presence of the full length product and four shorter versions thereof, were confirmed by this approach.

Thus, the four shorter transcripts were identical in all 49 transgenic moss lines and were present independent from the tissue type, i.e. whether juvenile or adult tissue was used. Accordingly, it was concluded that the different transcripts were not derived from a partial integration of the FIX expression construct into the *P. patens* genome.

In order to validate the presence of different FIX products also in transiently transfected moss cells, total RNA was extracted from transiently transfected cells (protoplasts) 14 days after transfection. After RT-PCR, it was confirmed again that four shorter variants next to the full length FIX product were present. Thus, exactly
as in the RT-PCR from stable lines, five FIX coding sequence variants (one full-length and four shorter versions thereof) were found in transiently transfected cells (Figure 1C).

Summarizing the above, the present example shows that the heterologous production of human FIX in (i) transgenic *P. patens* moss line Δxt/ft, namely either in (a) protonema cultures (juvenile cultures) or (b) gametophore cultures (adult cultures) as well as in (ii) transiently transfected *P. patens* protoplasts consistently resulted in the presence of five different FIX products, namely the full length product and four discrete shorter products thereof. The respective products had identical sequences across the moss lines used in the present example [(i) (a) and (b) as well as (ii)].

4.2. Example 2: Identification of splice motifs in the cDNA encoding FIX as potential explanation for the different discrete FIX-products

The aim of this example was to find an explanation on the nucleotide sequence-level for the presence of several discrete FIX-products.

The cDNA encoding FIX as used in example 1 corresponds to a coding sequence, where all introns have already been removed. Thus, as is derivable from the “c” in “cDNA”, only the coding sequence, i.e. the exons, without any introns are present. Therefore, a splicing process was thus far not considered for a cDNA that lacks introns since splicing usually takes place when removing the introns from the primary RNA transcript in order to arrive at a spliced RNA without any introns, i.e. in order to arrive at the cDNA.

The sequencing results of the different PCR products in the different moss lines discussed above in example 1 showed that the discrete four additional products are identical in their sequence and consistently present. The inventors therefore hypothesized that specific motifs in the human FIX cDNA might have been
recognized as alternative splice signals in *P. patens* with the result that some regions might have been treated as introns such that shorter versions than the full-length version are generated.

To further analyze the FIX transcripts produced in *P. patens*, the full-length transcript as well as its shorter versions were investigated in more detail on the sequence level. Potential splice junctions within the full-length FIX coding sequence were identified and checked against the lengths of the different products. It appeared from this analysis that splicing of the full length transcript from human cDNA in moss caused the production of four alternative FIX products as shown in Figure 2A:

i) deletion of a 44 amino acids (aa) long domain in the part coding for the heavy chain (Figure 2A.II);

ii) large deletion in the part coding for the light chain (115 aa) together with the deletion observed in the heavy chain (Figure 2A.III);

iii) almost complete deletion of the sequence coding for the light chain and parts of the activation peptide domain (157 aa) and the heavy chain (Figure 2A.IV);

iv) almost complete deletion of the sequence encoding the light chain, complete deletion of the activation peptide domain (186 aa) and the sequence representing 41 aa of the heavy chain (Figure 2A.V).

It is noted that the heavy chain harbors a serine protease domain, which is responsible for the activation of Factor X in the presence of Factor VIII, calcium and phospholipid surfaces following removal of the activation peptide. Any deletion in this domain will interfere with the activity of the FIX protein. The light chain consists of the Gla domain followed by two EGF domains. Any loss in these domains will adversely affect the properties of FIX. Therefore, any alterations induced by a potential alternative splicing process will result in the loss of FIX function in the coagulation cascade. Interestingly, a closer look at multiple sequence alignments of the FIX variants demonstrated that the potential alternative splicing process neither
caused a frameshift mutation nor an amino acid exchange at the respective splice site(s) (Figure 2B).

It was reported that the GT-AG splicing rule is applicable for almost all eukaryotic genes [46, 47], and especially *P. patens* accepts control elements originally optimized for mammalian expression systems without a need for adapting these elements [48]. When carrying out the above analysis on potential splice sites on a sequence level, four different motifs for acceptor and donor sites were identified. All potential FIX splice sites followed the GT-AG rule. Moreover, the data revealed that a CAGGT motif is present in the exon-intron junction, i.e. the donor site (Figure 3A). In the intron-exon junction at the acceptor site, the motif CAG is conserved in 3 out of 4 cases, but the adjacent two nucleotides do not follow any trend.

The splice sites are derivable from Figures 2A, 3A and 8 as follows:

- **Donor splice site 1** (SEQ ID NO: 1) with the sequence as given in Figure 3A and as shown by box 1 in Figure 8;
- **Acceptor splice site 2** (SEQ ID NO: 2) with the sequence as given in Figure 3A and as shown by box 2 in Figure 8;
- **Acceptor splice site 3** (SEQ ID NO: 3) with the sequence as given in Figure 3A and as shown by box 3 in Figure 8;
- **Donor splice site 4** (SEQ ID NO: 4) with the sequence as given in Figure 3A and as shown by box 4 in Figure 8;
- **Acceptor splice site 5** (SEQ ID NO: 5) with the sequence as given in Figure 3A and as shown by box 5 in Figure 8;
- **Acceptor splice site 6** (SEQ ID NO: 6) with the sequence as given in Figure 3A and as shown by box 6 in Figure 8; note that boxes 5 and 6 overlap such that box 6 is depicted above the sequences and box 5.

To check whether donor and acceptor sites and their neighboring nucleotides are conserved in *P. patens*, the genomic vicinity of all 87,533 annotated transcripts
corresponding to 32,926 protein-encoding genes of the current *P. patens* genome release v 3.3 were analyzed and the results are depicted in Figures 3B (this consensus sequence is SEQ ID NO: 7) and 3C (this consensus sequence is SEQ ID NO: 8). The experimental findings discussed above perfectly fit to the splice motifs retrieved from the analysis of all protein-coding transcripts.

Subsequently, the FIX coding sequence was analyzed with a moss splice site prediction tool using either Hidden Markov Models or Support Vector Machines on Cosmoss (http://cosmoss.org/ssp/). However, the tool largely failed to predict the experimentally identified splicing motifs. This might be due to the fact that the tool has been developed using only 368 donor and acceptor sites 13 years ago [51] and has not been updated since then.

Summarizing the above, the present example shows that the *P. patens* spliceosome obviously interprets parts of the FIX CDS as intronic sequences. One explanation for this might be the existence of cis exonic and intronic splicing enhancers [38]. Little is known about exonic and intronic regulatory sequences in plants [59-61]. The examples shows that - based on a sequence analysis on RNA level - (i) a potential alternative splicing process would perfectly explain the presence of four discrete shorter FIX products and that (ii) the GT-AG rule seems to apply to all potential FIX splice sites.

4.3 Example 3: Analysis of the FIX variants on protein level

The aim of the present example was to validate the findings obtained on RNA level on protein level by analyzing the produced moss FIX protein variants by mass spectrometry (MS).

The results on RNA level suggested that *P. patens* would produce four shorter FIX protein isoforms in addition to the full-length FIX. To validate this result, the culture
supernatants of transiently transfected cells were precipitated and analyzed with a polyclonal anti-FIX antibody via immunoblot after reduction and alkylation (Figure 4A). Full-length FIX was expected at a molecular mass at around 52 kDa and spliced FIX variants were calculated based on the alternatively spliced transcripts with molecular masses of 47, 34, 29, and 27 kDa, respectively. The bands observed on the immunoblot could be attributed to these isoforms together with a band at around 40 kDa (potentially a degradation product) and a band at around 70 kDa (potentially a dimer of one of the protein isoforms) (Figure 4A). Additionally, the other half of the supernatants was separated via SDS-PAGE, and, following Coomassie blue staining, the gel slice corresponding to the size of 25-70 kDa was divided into four pieces and prepared for MS determination of FIX-variant amino acid sequences. The MS analysis of the FIX-variants confirmed the presence of a unique FIX-variant peptide sequence. The full-length FIX sequence contains 44 amino acid between the peptides VVGGEDAKP (SEQ ID NO: 9) and GEHNIETEHTEQKR (SEQ ID NO:10) in the sequence coding for the heavy chain (Figure 4B). In case of splicing in this region, the 44 amino acids between these two sequences should be missing and the two sequences should be linked directly in order to arrive at the following sequence: VVGGEDAKPGEHNIETEHTEQKR (SEQ ID NO:11). This 24 amino acid long peptide was indeed identified in MS analysis after trypsin digestion, which confirms that splicing takes place from the predicted motif.

This example shows that (i) the full-length and the four shorter FIX variants thereof can also be identified on protein level and that (ii) at least one of the shorter FIX variants comprises a sequence that can only be generated if a sequence coding for 44 amino acids is spliced out of the full sequence. This splicing process is referred to in the following as “heterosplicing”.

4.4 Example 4: Sequence amendments on nucleotide level while maintaining the amino acid sequence and the impact of the amendments on heterosplicing
The aim of this example was to find out whether the undesired process of heterospliceing can be prevented by modifying the sequence of the FIX CDS (SEQ ID NO:21).

Two different versions of the FIX CDS were generated in a step-wise manner: first, the donor and acceptor sites were mutated arriving at aspFIX (SEQ ID NO:22), followed by modifying this aspFIX version further, namely by optimizing codon usage for the remaining codons arriving at optiFIX (SEQ ID NO:23).

Thus, to generate aspFIX, codons neighboring the splice junctions were modified such that the splice sites were destroyed while maintaining the encoded amino acids.

Furthermore, the possible actions of published *P. patens* miRNAs on this new sequence were analyzed. For this purpose, *A Plant Small RNA Target Analysis Server* (http://plantgrn.noble.org/v1_psRNATarget/?function=2 Analysis with Preloaded small RNAs/user-submitted transcripts) was used with the following search options:

<table>
<thead>
<tr>
<th>Maximum expectation</th>
<th>5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length for complementarity scoring</td>
<td>20</td>
</tr>
<tr>
<td># of top target genes for each small RNA</td>
<td>200</td>
</tr>
<tr>
<td>Target accessibility – allowed maximum energy to unpair the target site (UPE)</td>
<td>25</td>
</tr>
<tr>
<td>Flanking length around target site for target accessibility analysis</td>
<td>17 bp in upstream / 13 bp in downstream</td>
</tr>
<tr>
<td>Range of central mismatch leading to translational inhibition</td>
<td>9 – 11 nt</td>
</tr>
</tbody>
</table>

This analysis showed that ppt-miR1043-5p (CCUUCAACUCAUGCGUGCA, SEQ ID NO: 24) can bind to #347–#365 of FIX sequence. By changing CT TTGGA into
GTTCGGA, i.e. by amending to the two nucleotides shown in underlined and bold, the possible binding of this miRNA was prevented. The exact positions of these two nucleotides are positions 351 and 354 of the FIX sequence, see Figures 7 and 8. As for the splice junctions, the two nucleotides were amended while maintaining the encoded amino acids. Finally, although it did not seem from the analysis of the difference splice versions that this site was actually used, the consensus splice sequence GCAGGT at positions 711 to 717 was changed into GCAGAGTC, again while maintaining the encoded amino acids. The exact positions of these two nucleotides are positions 714 and 717 of the FIX sequence, see Figures 7 and 8.

_P. patens_ was transiently transfected with the resulting aspFIX plasmid (SEQ ID NO:25). On day 14, the cells were collected and RNA was isolated. RT-PCR and sequencing of the PCR products revealed that there were only two bands representing FIX variants: the full-length FIX coding sequence and a shorter variant at about 700 bp (Figure 5A). The presence of this new shorter variant of about 700 bp can be explained by the presence of new donor and acceptor sites in aspFIX (Figure 5B). The shorter variant of about 700 bp was different from the previously identified shorter variants and apparently misses the nucleotides encoding the light chain, AP and parts of the heavy chain (Figure 5C). Moreover, splicing would cause a frameshift mutation resulting in early stop codons, possibly resulting in RNA degradation via nonsense-mediated mRNA decay (NMD). However, as this transcript was detected by RT-PCR, it is highly likely that it is resistant to NMD. This would not be surprising since it is well known that not all nonsense mRNAs undergo NMD, which is e.g. reported for _P. patens_ [39] and _A. thaliana_ [52, 53]. The new splice sites in aspFIX are derivable from Figures 5A, 5B and 8 as follows:

- **Donor splice site 7** (SEQ ID NO:26) with the sequence as given in Figure 5B and as shown by box 7 in Figure 8;

- **Acceptor splice site 5′** (SEQ ID NO:27) with the sequence as given in Figure 5B and as shown by box 5′ in Figure 8;
It is noted that site 5' is identical to site 5 except for a single nucleotide exchange introduced in order to arrive at aspFIX. It is further noted that boxes 1 and 7 overlap such that box 7 is depicted above the sequences and box 1.

5

It is known that plant introns are different from animal introns in terms of UA- or U-richness, which is crucial for splicing efficiency [62]. It was shown before that the average GC content in \textit{P. patens} coding sequences is 50% [53]. The GC content of the human FIX CDS recognized as cryptic additional introns ranged from 38% to 40%. This remained the same after exclusively mutating the splice sites in order to arrive at aspFIX. However, even though mutations in the splicing motifs successfully decreased the number of transcripts from five (full length and four shorter variants) to two (full length and one shorter variant), the GC content of the cryptic alternative intron was still only 40%. Therefore, it was decided to create a fully codon-optimized plasmid (referred to as "optiFIX", see below) that should have an overall increased GC content.

To generate optiFIX, aspFIX was used as starting sequence and a complete codon optimization according to recent studies [45, 54] was carried out while maintaining the amendments carried out at codons neighboring the splice junctions. Of course, the amino acid sequence was not amended but the optiFIX sequence still encodes the wild-type full-length protein. \textit{P. patens} was transiently transfected with the resulting optiFIX plasmid (SEQ ID NO:28). On day 14, the cells were collected and RNA was isolated. RT-PCR and sequencing of the PCR products revealed that the full-length coding sequence of FIX was the only specific band. No further FIX variants were detected (Figure 5D). The sole presence of the full-length product was confirmed on protein level via immunoblot analysis (Figure 5E).

A comparison of the aspFIX and optiFIX constructs revealed that there was no change in the donor and acceptor sites of the newly emerged cryptic intron. There was, however, indeed an increase of the GC content within the cryptic intron
sequence in optiFIX. The GC contents of the sequences previously defined as introns in the FIX and aspFIX constructs increased to 46% - 50% in optiFIX.

It was therefore concluded that the increase of GC content introduced by the overall codon optimization together with mutating the splice sites prevented the heterosplicing, i.e. the mis-interpretation of any sequence as intron by the moss spliceosome.

In summary, this example demonstrates that (i) getting rid of the splice sites by modifying exclusively the codons next to the splice sites while maintaining the encoded amino acid sequence reduces the number of variants to the full-length FIX and a single spliced variant and that (ii) optimizing the further FIX coding sequence (at least partly to increase the GC-content) while maintaining the amendments at the splice sites and while overall maintaining the encoded amino acid sequence eliminates shorter variants of FIX and results in the production of only the full-length protein.

4.5 Example 5: The codon usage bias in *P. patens* and other heterologous expression systems

The aim of this example was to check the codon usage bias in different heterologous expression system.

The codon usage bias in *P. patens* was compared to biases of other heterologous production platforms. The bias in *P. patens* seems to be similar to codon usage biases in *Spodoptera frugiperda* (insect cells), *Homo sapiens* (HEK cells), *Cricetulus griseus* (CHO cells), and *Oryza sativa* (rice) but different to the codon usage bias in *Nicotiana tabacum* (tobacco) (Table 1).
Table 1 – The codon usage frequencies of under- and over-represented codons for eight amino acids in different protein expression systems. Numbers represent the frequency of codon per thousand. Data are from [http://www.kazusa.or.jp/codon/](http://www.kazusa.or.jp/codon/).

<table>
<thead>
<tr>
<th>Codon frequency in thousands</th>
<th>Physcomitrella patens</th>
<th>Spodoptera frugiperda</th>
<th>Homo sapiens (HEK cells)</th>
<th>Cricetulus grisens (CHO cells)</th>
<th>Oryza sativa</th>
<th>Nicotiana tabacum</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>under-represented TGT</td>
<td>7.0</td>
<td>8.5</td>
<td>10.6</td>
<td>9.1</td>
<td>6.2</td>
<td>9.8</td>
<td>Cys</td>
</tr>
<tr>
<td>over-represented TGC</td>
<td>9.8</td>
<td>13.2</td>
<td>12.6</td>
<td>10.3</td>
<td>12.4</td>
<td>7.2</td>
<td>Glu</td>
</tr>
<tr>
<td>under-represented GAA</td>
<td>24.1</td>
<td>27.6</td>
<td>29.0</td>
<td>28.4</td>
<td>21.6</td>
<td>36.0</td>
<td>Phe</td>
</tr>
<tr>
<td>over-represented GAG</td>
<td>37.4</td>
<td>33.1</td>
<td>39.6</td>
<td>41.1</td>
<td>38.6</td>
<td>29.4</td>
<td>His</td>
</tr>
<tr>
<td>under-represented TTT</td>
<td>17.2</td>
<td>10.1</td>
<td>17.6</td>
<td>19.6</td>
<td>13.1</td>
<td>25.1</td>
<td>Lys</td>
</tr>
<tr>
<td>over-represented TTC</td>
<td>23.6</td>
<td>27.5</td>
<td>20.3</td>
<td>22.0</td>
<td>22.4</td>
<td>18.0</td>
<td>Asn</td>
</tr>
<tr>
<td>under-represented CAT</td>
<td>11.1</td>
<td>8.7</td>
<td>10.9</td>
<td>10.2</td>
<td>11.3</td>
<td>13.4</td>
<td>Gln</td>
</tr>
<tr>
<td>over-represented CAC</td>
<td>11.6</td>
<td>15.6</td>
<td>15.1</td>
<td>12.9</td>
<td>13.8</td>
<td>8.7</td>
<td>Tyr</td>
</tr>
<tr>
<td>under-represented AAA</td>
<td>18.9</td>
<td>26.8</td>
<td>24.4</td>
<td>24.6</td>
<td>16.0</td>
<td>32.6</td>
<td></td>
</tr>
<tr>
<td>over-represented AAG</td>
<td>34.7</td>
<td>49.2</td>
<td>31.9</td>
<td>38.4</td>
<td>32.3</td>
<td>33.5</td>
<td></td>
</tr>
<tr>
<td>under-represented AAT</td>
<td>17.8</td>
<td>13.4</td>
<td>17.0</td>
<td>17.4</td>
<td>15.1</td>
<td>28.0</td>
<td></td>
</tr>
<tr>
<td>over-represented AAC</td>
<td>20.4</td>
<td>28.8</td>
<td>19.1</td>
<td>21.2</td>
<td>18.5</td>
<td>17.9</td>
<td></td>
</tr>
<tr>
<td>under-represented CAA</td>
<td>16.7</td>
<td>16.1</td>
<td>12.3</td>
<td>10.3</td>
<td>13.5</td>
<td>20.7</td>
<td></td>
</tr>
<tr>
<td>over-represented CAG</td>
<td>22.0</td>
<td>21.6</td>
<td>34.2</td>
<td>33.4</td>
<td>20.8</td>
<td>15.0</td>
<td></td>
</tr>
<tr>
<td>under-represented TAT</td>
<td>10.5</td>
<td>10.0</td>
<td>12.2</td>
<td>13.1</td>
<td>10.0</td>
<td>17.8</td>
<td></td>
</tr>
<tr>
<td>over-represented TAC</td>
<td>17.2</td>
<td>24.4</td>
<td>15.3</td>
<td>16.4</td>
<td>15.1</td>
<td>13.5</td>
<td></td>
</tr>
</tbody>
</table>

These results strongly suggest that the amendments in the CDS carried out if *P. patens* is used as expression system are applicable in other species/systems as well to overcome heterosplicing of heterologous cDNAs. For species that do not follow the codon choice of *P. patens* (such as *N. tabacum*) the GC content should still be
increased to prevent the heterosplcing by replacing codons ending with A or T to G or C but not necessarily with those disclosed in Table 1.

It is noted that *P. patens* has no codon preference for the remaining twelve amino acids shown in Table 2 below.

Table 2 – Codons without any preference in *P. patens*. The DNA codons given therein represent each amino acid as listed.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Codon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoleucine</td>
<td>ATT, ATC, ATA</td>
</tr>
<tr>
<td>Leucine</td>
<td>CTT, CTC, CTA, CTG, TTA, TTG</td>
</tr>
<tr>
<td>Valine</td>
<td>GTT, GTC, GTA, GTG</td>
</tr>
<tr>
<td>Methionine</td>
<td>ATG</td>
</tr>
<tr>
<td>Alanine</td>
<td>GCT, GCC, GCA, GCG</td>
</tr>
<tr>
<td>Glycine</td>
<td>GGT, GGC, GGA, GGG</td>
</tr>
<tr>
<td>Proline</td>
<td>CCT, CCC, CCA, CCG</td>
</tr>
<tr>
<td>Threonine</td>
<td>ACT, ACC, ACA, ACG</td>
</tr>
<tr>
<td>Serine</td>
<td>TCT, TCC, TCA, TCG, AGT, AGC</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>TGG</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>GAT, GAC</td>
</tr>
<tr>
<td>Arginine</td>
<td>CGT, CGC, CGA, CGG, AGA, AGG</td>
</tr>
</tbody>
</table>

4.6 Example 6: Comparison of the protein levels obtained when using FIX and optiFIX CDS *in vivo*

The aim of this example was to analyze the impact of using the optimized coding sequence vs. the original cDNA sequence on the protein expression inside the moss cells.

In order to check whether the protein amounts derived from the FIX sequence and the optiFIX sequence, respectively, show any difference in live moss cells, two fusion constructs, FIX-Citrine and optiFIX-Citrine, were engineered. The two
constructs were transiently transfected into moss protoplasts. Confocal laser scanning microscopy was performed on day 3 after transfection. A visual analysis of the images acquired for the proteins derived from the FIX sequence and the optiFIX sequence, respectively, reveals that the FIX-Citrine signal shows a sparse/diffuse localization pattern which might be due to intracellular mislocalization of aberrant protein isoforms (Figure 6A). In contrast, the optiFIX-Citrine exhibits an intricate network-like organization which indicates that the secretory pathway is occupied by this protein (Figure 6B). Furthermore, the optiFIX-Citrine signal intensity is greater than that of FIX-Citrine. This is reflected by the fact that, in order to attain similar signal intensities, different gain adjustments had to be used for acquisition of FIX-Citrine and optiFIX-Citrine samples (Table 3). In order to make the images quantitatively comparable, a normalization step was therefore performed, wherein the voxel intensities of each image were adjusted to 40% of detector gain. The comparison of mean voxel intensities calculated for FIX-Citrine and for optiFIX-Citrine (9.487 and 82.389 for 40% gain, respectively) shows an approximately 9-fold higher fluorescence intensity after the (i) splice site codon optimization and the (ii) additional codon optimization, which prevented undesired heterosplicing (Figure 6C). The raw data provided in Table 3 also includes quantitative information of original images without voxel normalization, the mean voxel intensities for the raw images, and the percentage of gains used to record each image.

Table 3 – Quantitative information of original images used in the analysis.

<table>
<thead>
<tr>
<th>Type</th>
<th>Gain % used for Acquisition</th>
<th>Mean Voxel Intensity</th>
<th>Normalized Mean Voxel Intensity for 40% Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIX_Raw_1</td>
<td>15.4</td>
<td>3.481</td>
<td>9.042</td>
</tr>
<tr>
<td>FIX_Raw_2</td>
<td>33</td>
<td>11.954</td>
<td>14.490</td>
</tr>
<tr>
<td>FIX_Raw_3</td>
<td>27</td>
<td>3.326</td>
<td>4.927</td>
</tr>
<tr>
<td>optiFIX_Raw_1</td>
<td>3.6</td>
<td>8.894</td>
<td>98.822</td>
</tr>
<tr>
<td>optiFIX_Raw_2</td>
<td>6</td>
<td>16.016</td>
<td>106.773</td>
</tr>
<tr>
<td>optiFIX_Raw_3</td>
<td>6</td>
<td>6.235</td>
<td>41.567</td>
</tr>
</tbody>
</table>
This example shows that the protein expression levels derived from the optiFIX sequence and the FIX sequence differ by a factor of about 9. Additionally, the localization of the proteins derived from the two different constructs was different, with a predominant localization of the proteins derived from optiFIX in the components of the secretory pathway of the cells.

4.7 Example 7: Materials and Methods

Design of the expression vectors pFIX, aspFIX, and optiFIX

The coding DNA sequence for human Factor IX (NCBI reference NM_000133) without the native signal peptide was synthesized as N-terminal translational fusion to the signal-peptide coding sequence of PpAP1 (Pp3c5_19520V3.1) [55, 56] by GeneArt (Thermo Fisher Scientific, Waltham, MA, USA) (the sequence of the FIX CDS is given in SEQ ID NO:21) and cloned into the final expression vector pFIX, where the expression is driven by the promoter of the PpActin5 gene (SEQ ID NO:29) [20, 57], further comprising a CaMV 35S terminator (SEQ ID NO:30). The sequence of the resulting plasmid is given in SEQ ID NO:31. Nucleotides involved in splicing were replaced as explained above without a change of the FIX aa sequence in the aspFIX plasmid, which was synthesized also by GeneArt. For codon optimization, two sources – codon usage bias calculated by Hiss et al. (2017) and Kazusa codon usage database (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=145481) were used. Under-represented codons were manually replaced with the over-represented alternative codons (Table 1, P. patens) and the optiFIX construct was synthesized (Figure 7 for a multiple sequence alignment of FIX (SEQ ID NO:21), aspFIX (SEQ ID NO:22), and optiFIX (SEQ ID NO:23)).

Transfection and screening of transgenic plants
The *Physcomitrella patens* (Hedw.) B.S. Δxt/ft moss line, a double knockout for the α1,3 fucosyltransferase and the β1,2 xylosyltransferase genes [20] (IMSC accession number 40828), was cultured in liquid Knop medium (pH 4.5) one week prior to transfection as described previously [63] to obtain cells (protoplasts). Protoplast transfection was performed as described earlier [64] using 50 μg of linearized pFIX DNA per transfection. In addition to the generation of transgenic lines, mossFIX was produced transiently by an upscaled transfection protocol based on the proportion of 0.7-1 μg DNA of pFIX for 1x10^5 cells. Stable transfectants were selected on solidified Knop medium containing 25 μg/mL hygromycin as described previously [63]. Cells transiently expressing mossFIX were grown in special regeneration medium and harvested for RNA and protein analyses two weeks after transfection. To check whether splicing of FIX was prevented on RNA level, WT cells were also transfected with the aspFIX and optiFIX construct, respectively and harvested for RNA isolation two weeks after transfection.

*Plant cell culture*

After selection on solidified Knop medium, transgenic lines were cultured under standard conditions in liquid Knop medium according to [9, 63]. They were subcultured periodically by disrupting the tissue with the use of an ultraturrax (IKA, Staufen, Germany) at a rotational speed of 16,000 – 18,000/min for 1 min and transferred to fresh medium.

*Molecular characterization of transgenic lines*

Following the hygromycin selection process, stable lines were characterized by the presence of a FIX transcript. For transgene expression analysis via reverse transcription (RT)-PCR, 50-100 mg (fresh weight) moss material from protonema cultures were disrupted with the use of a stainless steel ball in TissueLyser II (Qiagen, Hilden, Germany) for 2 min with an impulse frequency set to 30. The
material was resuspended in 1 mL TRIzol® Reagent (Thermo Fisher Scientific), and RNA was isolated according to the manufacturer’s instructions. For analysis of gametophores, 5-6 gametophores were collected and RNA isolation was performed as described above. Six micrograms of DNase-treated RNA were used for first strand synthesis with Superscript Reverse Transcriptase III (Thermo Fisher Scientific). The quality of cDNA was controlled with a standard PCR with the primers c45for (5’-GGTTGCTCATGGTGTGCG-3’ - SEQ ID NO:32) and c45rev (5’-GAGGTTAAGCTCTGC-3’ - SEQ ID NO:33) corresponding to the gene coding for the ribosomal protein L21. Transgene expression was verified with the primers FIXfwdB (5’-GGAGTGTTCGATTGGGCTTCT-3’ – SEQ ID NO:34) and FIXrevB (5’-TGTTGGTGTTGGGTTAAGTT-3’ – SEQ ID NO:35) that amplifies almost the full-length FIX CDS. PCR reactions were performed with Phusion high-fidelity DNA polymerase (Thermo Fisher Scientific). The PCR products were examined by standard agarose gel electrophoresis with visualization of DNA by ethidium bromide fluorescence.

Molecular characterization of transiently transfected cells

Following the transfection, cells were grown in regeneration medium for two weeks. Transiently transfected cells (nearly 3.6 million cells for FIX and aspFIX; 7.2 million cells for optiFIX) were used for transgene expression analysis via RT-PCR. RNA was isolated using the TRIzol® Reagent and RNA was purified through spin columns of Direct-zol™ RNA MicroPrep kit (Zymo Research, Irvine, California, USA) according to the manufacturer’s instructions. After DNase-treatment and first-strand synthesis (for the cDNA synthesis: 0.4 µg RNA was used for FIX and aspFIX; 2 µg RNA was used for optiFIX), the quality of cDNA was controlled with primers c45for and c45rev. Transgene expression was verified again with primers FIXfwdB and FIXrevB.

Analysis of FIX cDNA products
Amplified products were excised from the gel and purified using QIAEX II Gel Extraction Kit (Qiagen). Sequencing was done by GATC Biotech AG (Sanger sequencing using the light run sequencing service). In the sequencing reactions, in addition to FIXfwdB and FIXrevB, FIX6deepseq1F (ATGGGGGCATCGAGGAGTGTCCGAT – SEQ ID NO:36), FIX6deepseq1R (TTCCCCAGCCACTTACATAGCCAGA – SEQ ID NO:37), and FIXseq6 (CACCAACAACCCGAGTGAAG – SEQ ID NO:38) were used. The remaining purified products were separately cloned into the pJET vector using the CloneJET PCR Cloning Kit (Thermo Fisher Scientific) according to the manufacturer’s instructions. Clones with aberrant insert sizes were selected for plasmid isolation with the GeneJet Plasmid Miniprep kit (Thermo Fisher Scientific). Sequencing was done as described above.

15 Splicing motif search

The most recent release of the Physcomitrella patens v3.3 genome annotation [5], which is available from https://phytozome.jgi.doe.gov/, was used. The splice sites of all 87,533 annotated transcripts from all 32,926 protein-coding genes were used as anchors to extract surrounding genomic sequences with a fixed window size of ± 9 bp. Consensus sequence logos of all extracted donor and acceptor sites were created using R and the package ggsseqlogo [49,65].

Codon usage bias among different species

25 Synonymous codon usage biases in P. patens, S. frugiperda, H. sapiens, C. griseus, O. sativa, and N. tabacum were obtained from Codon Usage Database (https://www.kazusa.or.jp/codon/) and biased codons in moss were compared with the codon usage pattern of other species.
Confocal microscopy imaging and image analysis

Image acquisition: All images were taken with a Leica TCS SP8 microscope (Leica Microsystems, Wetzlar, Germany) using HCX PL APO 63x/1.40 oil objective with a zoom factor of 5. The voxel size was 0.072 μm on the X-Y dimensions and 0.240 μm on the Z dimension. The pinhole was adjusted to 1 Air Unit (95.5 μm). For the excitation of Citrine Argon laser was applied at 4% (with an output power of 3.97 W) with an excitation wavelength of 514 nm. The detection range for Citrine fluorescence was set to 532-569 nm in the HyD detector. Z-stacks of 6 cells (3 replicates for FIX and 3 for OptiFIX) were acquired. For image analysis the FIJI software was used [66]. For each Z-stack the brightest Z-slice was determined, and then by taking this brightest slice as the midpoint, 9 slices were extracted for subsequent analysis. The voxel intensities of all images were normalized to a detector gain of 40%. Based on these normalized values the mean voxel intensity was determined for every image. The bar plot was generated for mean voxel intensities of 3 biological replicates for each group. For the 3D reconstruction of Z-stacks blend rendering was performed by using IMARIS 9.2.0 software (Bitplane AG, Zurich, Switzerland).

Recombinant protein extraction and detection

Culture supernatant was precipitated with a previously described method using chloroform/methanol [67]. After air drying, the precipitate was resuspended in 50 mM HEPES, 2% SDS (pH 7.5) and disulfide bonds within proteins were reduced with 25 mM dithiothreitol (DTT), followed by alkylation of the thiol groups with 60 mM iodoacetamide (IAA). For Western blot analyses, 7.5% SDS-PAGE (Ready Gel® Tris-HCl Precast Gels, BioRad, CA, USA) was run at 100 V for 1h 30min and blotted to polyvinylidene fluoride (PVDF) membranes (Amersham Hybond P 0.45 μm pore size PVDF blotting membrane, GE Healthcare) in a Trans-Blot® SD Semi-Dry Transfer Cell (BioRad) for 2 h with 1.8 mA/cm2 membrane. The membrane was
blocked for 1 h at room temperature in TBS containing 4% Amersham ECL Prime Blocking Reagent (GE Healthcare) and 0.1% Tween 20. It was then incubated with polyclonal anti-factor IX antibody (F0652, SIGMA) (1:3000) overnight at 4°C. Afterwards, the membrane was washed three times with TBS containing 0.1% Tween 20 and incubated with horseradish peroxidase-linked anti-rabbit IgG (NA934, GE Healthcare) at a dilution of 1:10,000 for 1 h. The blot was washed again and proteins were detected using the Amersham ECL Primer Western Blotting Detection Reagent (GE Healthcare) according to the manufacturer’s instructions.

Sample preparation and MS analysis

The reduced and alkylated protein extracts were subjected to SDS polyacrylamide gel electrophoresis. Appropriate gel bands showing FIX fractions were excised and processed for in-gel digestion: Following the destaining of the gel bands with 30% acetonitrile (ACN), they were shrinked with 100% ACN and dried completely in a vacuum concentrator. Tryptic in-gel digestion of gel slices was performed overnight at 37°C in 50 mM ammonium bicarbonate using 0.1 μg trypsin (Promega) per gel band. Afterwards, peptides were extracted from the gel with 5% formic acid. Analyses were performed using the UltiMate 3000 RSLCnano system (Dionex LC Packings/Thermo Fisher Scientific) coupled online to a QExactive Plus instrument (Thermo Fisher Scientific). The UHPLC systems was equipped with a C18-precolumn (Ø: 0.3 mm, 5 mm; PepMap, Thermo Fisher Scientific) and an Acclaim® PepMap analytical column (ID: 75 m, 500 mm, 2 m, 100 Å, Dionex LC Packings/Thermo Fisher Scientific).

MS analyses were performed using a binary solvent system consisting of 0.1% formic acid (FA, solvent A, “A”) and 0.1% FA/86% ACN (solvent B, “B”). Samples were washed and pre-concentrated on a C18-precolumn with 0.1% TFA for 5 min before switching the column in line with the analytical column. Peptide separation was performed applying a 45-min gradient at a flow rate of 250 nl/min. Peptide
samples were eluted with a gradient of 4–40% B in 30 min and 40–95% B in 5 min. After each gradient, the analytical column was washed with 95% B for 5 min and re-equilibrated for 15 min with 4% B. The MS instrument was externally calibrated using standard compounds and equipped with a nanoelectrospray ion source and distal coated SilicaTips (FS360–20-10-D, New Objective, Woburn, MA) and MS/MS analyses were performed on multiply charged peptide ions. The instrument was operated in the data-dependent mode to automatically switch between MS (max. of 1x10 ions) and MS/MS. Each MS scan was followed by a maximum of 12 MS/MS scans using HCD with a normalized collision energy of 35%. The mass range for MS was m/z = 375–1,700 and resolution was set to 70,000. MS parameters were as follows: spray voltage 1.5 kV; ion transfer tube temperature 200°C. Raw data were analyzed using Mascot Distiller V2.5.1.0 (Matrix Science, USA) and the peak lists were searched with Mascot V2.6.0 against an in-house database containing all P. patens V1.6 protein models [4] as well as human FIX (complete and alternative ones based on splicing).

4.8 Example 8: Optimization of the expression of the CMP-sialic acid transporter (CST) in *P. patens*

The aim of the present example was to apply the optimization steps carried out for FIX to a different protein and to check whether this also results in an increased production of the protein of interest. The methods as described in example 7 were used accordingly.

In the present example, the protein of interest was the CMP-sialic acid transporter (CST). The cDNA from the *Mus musculus* CST without introns and with a length of 1008 bp was used (NCBI Reference Sequence: NM_011895.3; SEQ ID NO:44). The cDNA was integrated into a suitable expression construct, which was transfected into the moss *P. patens*. CST, a highly hydrophobic type III membrane protein, is
responsible for the transport of CMP-sialic acid from the cell nucleus into the Golgi apparatus.

RT-PCR analysis of the expression products was carried out. This analysis showed that there was – as for FIX - heterosplicing of the CST mRNA in *P. patens*.

Accordingly, the steps as carried out for FIX were also carried out in the present example: (i) splice sites were modified such that the splice sites were destroyed while maintaining the encoded amino acid sequence and (ii) the resulting sequence was optimized further in that under-represented codons were replaced with over-represented codons. In sum, 69 of 336 codons were changed and the GC% content increased from 44.2% to 50.3% (the resulting cDNA sequence is SEQ ID NO:45) The amendments to the original CST cDNA ("CST") can be derived from Figure 9, where CST is shown in alignment with the modified sequence ("optiCST").

In order to analyze whether the protein amounts resulting from the expression using CST as cDNA and optiCST as cDNA are different in *P. patens*, two fusion constructs were cloned, namely CST-Citrine and optiCST-Citrine. As in example 6, the expression can then be compared inside the moss cells after transient transfection of the constructs into moss protoplasts. To this aim, fluorescent microscopy was carried out on day 3 after transfection. A visual analysis of the images acquired for CST-Citrine and optiCST-Citrine revealed that the optiCST-Citrine signal intensity is higher than that of CST-Citrine, see Figure 10.

4.9 Example 9: Optimization of the expression of the complement factor H (FH) in *P. patens*

The aim of the present example was to apply the optimization steps carried out for FIX and CST to yet a different protein and to check whether this also results in an
increased production of the protein of interest. The methods as described in example 7 were used accordingly.

In the present example, the protein of interest was the human Complement factor H (FH). The cDNA from human FH without introns and with a length of 3639 bp was used, as for FIX fused to a sequence encoding a plant signal peptide, in the present example a sequence encoding the signal sequence of the H1-gene from Thuja occidentalis (NCBI Reference Sequence of FH: Gene ID: 3075; SEQ ID NO:46, it is noted that this sequence includes the afore-mentioned signal peptide-encoding sequence). The cDNA was integrated into a suitable expression construct, which was transfected into the moss P. patens. FH is a soluble glycoprotein that has a unique role in the complement cascade, as it is the main negative regulator of the alternative pathway. Deficiencies in FH functionality lead to severe diseases, like atypical hemolytic uremic syndrome (aHUS), age-related macular degeneration (AMD) or C3 glomerulopathies (C3G). It is a single-chain, 155 kDa glycoprotein continuously expressed at a basal level in the liver, leading to typically 200-300 micrograms protein per milliliter blood. FH is composed of 20 globular domains. These domains are termed short consensus repeats (SCRs). These autonomously folded globular domains are stabilized via two internal disulfide bonds, respectively, leading to 2x20=40 internal disulfide bridges per protein molecule. FH acts in the blood plasma as well as on surfaces of host cells.

RT-PCR analysis of the expression products was carried out. This analysis showed that there was – as for FIX and CST- heterosplicing of the FH mRNA in P. patens. Thus, the RT-PCR analysis revealed that human FH cDNA generates two transcript variants in P. patens: the full-length FH transcript of 3639 bp and a 1215 bp heterospliced variant resulting in an FH isoform with a molecular mass of 45.6 kDa.

Accordingly, the steps as carried out for FIX and CST were also carried out in the present example: (i) splice sites were modified such that the splice sites were
destroyed while maintaining the encoded amino acid sequence and (ii) the resulting sequence was optimized further in that under-represented codons were replaced with over-represented codons. In sum, 386 of 1213 codons were changed and the GC% content increased from 33.9% to 49.8% (the resulting cDNA sequence is SEQ ID NO:47). The amendments to the original FH cDNA ("FH") can be derived from Figure 11, where FH is shown in alignment with the modified sequence ("optiFH").

In order to analyze whether the protein amounts resulting from the expression using FH as cDNA and optiFH as cDNA are different in *P. patens*, two fusion constructs were cloned, namely FH-Citrine and optiFH-Citrine. As in example 6, the expression can then be compared inside the moss cells after transient transfection of the constructs into moss protoplasts. To this aim, fluorescent microscopy was carried out on day 3 after transfection. A visual analysis of the images acquired for FH-Citrine and optiFH-Citrine revealed that FH-Citrine signal intensity is below the detection limit. On the other hand, optiFH-Citrine signal intensity is high, see Figure 12.

References


17. Shen JS, Busch A, Day TS, Meng XL, Yu CI, Dabrowska-Schlepp P, Fode B,


53. Frank W, Decker EL, Reski R (2005) Molecular tools to study *Physcomitrella*


CLAIMS

1. An \textit{in vitro} method of obtaining a cDNA for recombinant heterologous expression in a cell, wherein said method comprises the following steps:
   a) Providing a cDNA;
   b) Identifying nucleotides of consensus splice sites in the cDNA provided in step a);
   c) Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;
   d) Identifying under-represented codons in cDNA NO:1 obtained in step c);
   e) Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2,
wherein cDNA NO:2 corresponds to the cDNA for recombinant expression in a cell; wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression; and wherein the cDNA is mammalian cDNA.

2. An \textit{in vitro} method of optimizing a cDNA for recombinant heterologous expression in a cell, wherein said method comprises the following steps:
   a. Providing a cDNA;
   b. Identifying nucleotides of consensus splice sites in the cDNA provided in step a);
   c. Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;
d. Identifying under-represented codons in cDNA NO:1 obtained in step c);

e. Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2,

wherein cDNA NO:2 corresponds to a cDNA optimized for recombinant expression in a cell; wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression; and wherein the cDNA is mammalian cDNA.

3. An in vitro method of producing a protein by recombinant heterologous expression of a cDNA in a cell, wherein said method comprises the following steps:

a. Providing a cDNA;

b. Identifying nucleotides of consensus splice sites in the cDNA provided in step a);

c. Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;

d. Identifying under-represented codons in cDNA NO:1 obtained in step e);

e. Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2;

f. Recombinantly expressing cDNA NO:2 obtained in step e) in the cell, thereby producing the protein wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression; and wherein the cDNA is mammalian cDNA.
4. An *in vitro* method of optimizing the recombinant heterologous expression of a cDNA in a cell, wherein said method comprises the following steps:
   a. Providing a cDNA;
   b. Identifying nucleotides of consensus splice sites in the cDNA provided in step a);
   c. Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;
   d. Identifying under-represented codons in cDNA NO:1 obtained in step c);
   e. Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2;
   f. Using cDNA NO:2 obtained in step e) for recombinant expression in a cell, thereby optimizing the recombinant expression of a cDNA in a cell wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression; and wherein the cDNA is mammalian cDNA.

5. An *in vitro* method of increasing the fraction of the full-length variant of a protein produced by recombinant heterologous expression of a cDNA in a cell, wherein said method comprises the following steps:
   a. Providing a cDNA encoding the full-length variant of a protein;
   b. Identifying nucleotides of consensus splice sites in the cDNA provided in step a);
   c. Amending at least one of the nucleotides identified in step b) such that (i) at least one consensus splice site is destroyed while (ii) the
encoded amino acid sequence is maintained, thereby arriving at cDNA NO:1;

d. Identifying under-represented codons in cDNA NO:1 obtained in step c);

e. Replacing the under-represented codons identified in step d) with over-represented codons, thereby arriving at cDNA NO:2;

f. Using cDNA NO:2 obtained in step e) for recombinant expression in a cell, thereby obtaining an increased fraction of the full-length variant of the protein wherein the identification of under-represented codons in step d) and the replacement with over-represented codons in step e) is based on the codon usage of the cell used for recombinant expression; and wherein the cDNA is mammalian cDNA.

6. The method according to any one of claims claim 3 to 5, wherein said method comprises a step prior to step b), namely the step of recombinantly expressing the cDNA provided in step a) in the cell and determining whether the expression results in splice variants of the encoded protein.

7. The method according to any one of the preceding claims, wherein the consensus splice sites are donor and acceptor splice sites.

8. The method according to claim 7, wherein the donor and acceptor splice sites follow the GT-AG Rule.

9. The method according to any one of the preceding claims, wherein the mammalian cDNA is selected from the group consisting of human cDNA, rat cDNA and mouse cDNA.
10. The method according to any one of the preceding claims, wherein said cell is selected from the group consisting of a plant cell, a mammalian cell and an insect cell.

11. The method according to any one of the preceding claims, wherein said cell is a plant cell, preferably a *P. patens* cell.

12. The method according to any one of the preceding claims, wherein said mammalian cDNA is human cDNA, rat cDNA or mouse cDNA, preferably human cDNA; wherein said cell is a plant cell, preferably a *P. patens* cell; wherein the donor and acceptor splice sites follow the GT-AG Rule; and wherein the under-represented and over-represented codons are derived from the codon usage in said plant cell, preferably in the *P. patens* cell.

13. Use of a cDNA obtained according to the method of claim 1 for recombinant heterologous expression in a cell.

14. Use of a cDNA optimized according to the method of claim 2 for recombinant heterologous expression in a cell.

15. Use of the method according to claim 3 in the generation of a host cell for recombinant heterologous expression of a mammalian cDNA, wherein said host cell is capable of carrying out at least one posttranslational modification found in the mammal.

16. A protein produced by the *in vitro* method according to claim 3.
Figure 6

A

B

C

Mean Voxel Intensity

FIX-Citrine  optIFIX-Citrine

0  20  40  60  80  100  120
Figure 9 – continued

| CST     | 781 | TCAGTGCTTGTTGAAAGTATAACAGACACATGAAAGGCTCTCTGCTGCCGAGGCAATT |
| optiCST | 781 | TCAGTGCTTGTTGAAAGTATAACAGACACATGAAAGGCTCTCTGCTGCCGAGGCAATT |
| consensus | 781 | ******************************************** |

| CST     | 841 | GTCTCTTCTACCATGGCTTCAGTCTCACTGTTGATACAGATGATAACACTTTTGATTGCA |
| optiCST | 841 | GTCTCTTCTACCATGGCTTCAGTCTCACTGTTGATACAGATGATAACACTTTTGATTGCA |
| consensus | 841 | *** ** *** ** *** ****** ****** *** |

| CST     | 901 | CTGGAAGCTCTTCTGTGTGGTCTGATCCATATATCTCTATGTTGATTCCAGACAGGACT |
| optiCST | 901 | CTGGAAGCTCTTCTGTGTGGTCTGATCCATATATCTCTATGTTGATTCCAGACAGGACT |
| consensus | 901 | ************ ****** ****** ****** ****** *** |

| CST     | 961 | ACATCCATTCAAAGAGCAAGCTCTCAAAGAGCAACATTGATGTCATGTTG |
| optiCST | 961 | ACATCCATTCAAAGAGCAAGCTCTCAAAGAGCAACATTGATGTCATGTTG |
| consensus | 961 | ************ ****** ****** ****** ****** ****** *** |

Figure 10

A

B

20 μm

20 μm
Figure 11 – continued

FH  GGC7GGAATGCCCTGCTCCAGGATGATCCCTGAAACCTTGAGTTTATCCGAGACATTAAACAT 1020
optiFH GGC7GGAATGCCCTGCTCCAGGATGATCCCTGAAACCTTGAGTTTATCCGAGACATTAAACAT 1020
******** ************** ******** **** ******* **

FH  GGAAGTCTATATATATGAGAATATATGCGTAGGCACATTTTCTCCGATGATGAAATAAT 1080
optiFH GGAAGTCTATATATATGAGAATATATGCGTAGGCACATTTTCTCCGATGATGAAATAAT 1080
******** ******* ******** *** ** ******* **

FH  TACTCTTATTACGTGATGACAACTTATTTTGAGACTGCTCCGAGGAAATGTTACTGCGGATACATT 1140
optiFH TACTCTTATTACGTGATGACAACTTATTTTGAGACTGCTCCGAGGAAATGTTACTGCGGATACATT 1140
******** ** ******** ** ** ** ******** **

FH  CATGTCAGCAACAGAGATGAGGATTCTTGCCGACGAGTACATGCTTCGGAATAATTTTCT 1200
optiFH CATGTCAGCAACAGAGATGAGGATTCTTGCCGACGAGTACATGCTTCGGAATAATTTTCT 1200
** ******** ******** ** ** ** ******** **

FH  TATTTGGGAAATATGAGATATCTGATATCTCGCCAGGATTGCAAGAAGTTTGTACGAGGTAATAATTCTATT 1260
optiFH TATTTGGGAAATATGAGATATCTGATATCTCGCCAGGATTGCAAGAAGTTTGTACGAGGTAATAATTCTATT 1260
** ******** ** ******** ** ** ** ******** **

FH  GACTCTTGCTGTCATTGCTCTCCGAGTCCTACGCTCATCCTCAAAGCCGACGACAGATTATCACTGATG 1320
optiFH GACTCTTGCTGTCATTGCTCTCCGAGTCCTACGCTCATCCTCAAAGCCGACGACAGATTATCACTGATG 1320
******** ** ******** ** ******** ** ******** **

FH  GAGAGTCGGTCTGCCCTCCGACCCGCCGCCATCCGCTACATTGCCCAAAAGCCGCAGAAGACTATCACTGATG 1380
optiFH GAGAGTCGGTCTGCCCTCCGACCCGCCGCCATCCGCTACATTGCCCAAAAGCCGCAGAAGACTATCACTGATG 1380
******** ** ******** ** ******** ** ******** **

FH  ATAGATTTGGAAGATGTTTTATTTTGATCTGAAATCTGACCATATATGATGCTTTCT 1440
optiFH ATAGATTTGGAAGATGTTTTATTTTGATCTGAAATCTGACCATATATGATGCTTTCT 1440
** ******** ** ******** ** ******** **

FH  GCACAAATATCTGACAAACTAGGATATGTAACAGCAGATGTTGAAAACATCTGCAAAATATTATT 1500
optiFH GCACAAATATCTGACAAACTAGGATATGTAACAGCAGATGTTGAAAACATCTGCAAAATATTATT 1500
** ** ** ** ******** ** ******** **

FH  ACAAGGTTGGGGAGAAGGATGAGGATGCAGCCTAACCCACATGCGAATTTAATCTGTTGATATCCCA 1560
optiFH ACAAGGTTGGGGAGAAGGATGAGGATGCAGCCTAACCCACATGCGAATTTAATCTGTTGATATCCCA 1560
******** ** ******** ** ******** ** ******** **

FH  GTATTCTGTTATGAGAAGATGTTTTATGGAAGAACTATCCGAGAGATGTTTAAAGCTGAGTAACACTTG 1620
optiFH GTATTCTGTTATGAGAAGATGTTTTATGGAAGAACTATCCGAGAGATGTTTAAAGCTGAGTAACACTTG 1620
******** ** ******** ** ******** ** ******** **

FH  GACCTATTGAAAGGCTGACAGGATGATTCTCCGAGGAACTCAGGAGGAGCAAGATGTTATATG 1680
optiFH GACCTATTGAAAGGCTGACAGGATGATTCTCCGAGGAACTCAGGAGGAGCAAGATGTTATATG 1680
******** ** ** ** ******** ** ******** **

FH  TGCGGTTGAAACGTTGGTTGCTGGTATTTATCGCAATTGGTATGAAAGAGAGAATGCGAATCTTCCCT 1740
optiFH TGCGGTTGAAACGTTGGTTGCTGGTATTTATCGCAATTGGTATGAAAGAGAGAATGCGAATCTTCCCT 1740
** ******** ** ******** ** ******** **

FH  AAAATTGATGTACAGGACTTTGATGTCAGATGCCAAAGAAAGAGACAGGATATAGGTGGGAGAGGTTG 1800
optiFH AAAATTGATGTACAGGACTTTGATGTCAGATGCCAAAGAAAGAGACAGGATATAGGTGGGAGAGGTTG 1800
** ** ******** ** ******** ** ******** **

FH  TGGCGATCTTCTGCTGCAAACTCCGAGGATTATATGATGTTGGGACTCATTGCTCTGAGGAGGAGGAGATG 1860
optiFH TGGCGATCTTCTGCTGCAAACTCCGAGGATTATATGATGTTGGGACTCATTGCTCTGAGGAGGAGGAGATG 1860
******** ** ******** ** ******** ** ******** **
Figure 11 – continued

FH
ACATTCGAGAACATCTTGGGATGGGGAAAACCTGGAGTAGTATCGACACTTTTGCTGCAAAAAGATAA 3720
optiFH
ACATTCGAGAACATCTTGGGATGGGGAGCTGGGAAGCTGGAGTAGTATCGACACTTTTGCTGCAAAAAGATAA 3720
********************************************** ******

Figure 12

A

B

3 nm

3 nm
**INTERNATIONAL SEARCH REPORT**

**A. CLASSIFICATION OF SUBJECT MATTER**

INV. C12N15/00  C12N15/10  C12N15/82  C12N9/64

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

C12N  C40B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data, INSPEC, BIOSIS, EMBASE

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>MARK A. JACKSON ET AL: &quot;Design rules for efficient transgene expression in plants&quot;, PLANT BIOTECHNOLOGY JOURNAL, vol. 12, no. 7, 23 May 2014 (2014-05-23), pages 925-933, XP055594229, GB ISSN: 1467-7644, DOI: 10.1111/pbi.12197 abstract; figures 2-5; tables 1,2 page 928 - page 931</td>
<td>1-16</td>
</tr>
</tbody>
</table>

Further documents are listed in the continuation of Box C. See patent family annex.

**Date of the actual completion of the international search**

3 April 2020

**Date of mailing of the international search report**

15/04/2020

**Name and mailing address of the ISA/**

European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-3040, Fax: (+31-70) 340-3016

**Authorized officer**

Tilkorn, A
<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>HABIBI PEYMAN ET AL: &quot;Optimization of inside and outside factors to improve recombinant protein yield in plant&quot;, PLANT CELL, TISSUE AND ORGAN CULTURE, SPRINGER, NL, vol. 130, no. 3, 2 June 2017 (2017-06-02), pages 449-467, XPO36300985, ISSN: 0167-6857, DOI: 10.1007/S11240-017-1240-5 [retrieved on 2017-06-02] page 454 - page 456; figure 1; table 1</td>
<td>1-16</td>
</tr>
<tr>
<td>Patent document cited in search report</td>
<td>Publication date</td>
<td>Patent family member(s)</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>-----------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 2016004113 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 2312291 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP 2002517180 A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 6642028 B1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US 2003224508 A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WO 9929848 A1</td>
</tr>
</tbody>
</table>