

Nucleic acids encoding repetitive amino acid sequences rich in proline and alanine residues that have low repetitive nucleotide sequences

The present invention relates to a nucleic acid molecule comprising a low repetitive nucleotide sequence encoding a proline/alanine-rich amino acid repeat sequence. The encoded polypeptide comprises a repetitive amino acid sequence that forms a random coil. The nucleic acid molecule comprising said low repetitive nucleotide sequences can further comprise a nucleotide sequence encoding a biologically or pharmacologically active protein. Further, the present invention provides for selection means and methods to identify said nucleic acid molecule comprising said low repetitive nucleotide sequence. The present invention also relates to a method for preparing said nucleic acid molecule(s). Also provided herein are methods for preparing the encoded polypeptide or drug conjugate(s) with the encoded polypeptide using the herein provided nucleic acid molecules. The drug conjugate may comprise a biologically or pharmacologically active protein or a small molecule drug. Also provided herein are vectors and hosts comprising such nucleic acid molecules.

Polypeptides that form random coils are known in the prior art. For example, WO 2008/155134 discloses proteins comprising an amino acid sequence of at least about 100 amino acid residues and consisting of proline, alanine and serine (PAS) residues. The amino acid sequence forming the random coil conformation can comprise a plurality of amino acid repeats. These repeats may consist of at least 3 to 30 or more amino acid residues. WO 2011/144756 discloses polypeptides comprising repetitive amino acid sequences consisting solely of proline and alanine (PA) residues. These polypeptides also form random coils and consist of at least 50 proline and alanine residues. WO 2015/132004 discloses a recombinant clostridial neurotoxin comprising a random coil domain consisting of PAS. US 2006/0252120 A1 discloses hydroxyproline-rich glycoproteins, which contain segments encoded as proline-rich glycomodules with the amino acid sequence motif $[(AP)_5]_n$. Also naturally occurring polypeptides encompass sequences rich in proline and alanine, such as the very large tegument protein of the *Macacine herpesvirus 1* gene published under the gene bank accession number (AAP41454.1). Methods for codon optimization are disclosed in WO 2007/142954.

Repetitive prior art polypeptides, like PAS or PA sequences, are typically encoded by corresponding repetitive nucleic acids. Accordingly, the prior art nucleic acids reflect the repetitive structure of the amino acid sequences they encode also in their nucleotide sequence.

- 2 -

Thus, the prior art nucleic acids are highly repetitive on their sequence level. The repetitiveness of prior art nucleic acids can lead to caveats such as partial genetic instability, in particular when encoding long PAS or PA sequences, for example 300 residues or longer.

The technical problem underlying the present invention is therefore the provision of means and methods for the convenient and reliable preparation of polypeptides containing amino acid repeats, in particular polypeptides containing amino acid repeats that consist of proline, alanine and, optionally, serine.

The technical problem is solved by provision of the embodiments provided herein below and as characterized in the appended claims.

The present invention relates to the following items:

1. A nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein the nucleotide sequence of said nucleic acid has a length of at least 300 nucleotides, wherein said nucleotide sequence has a Nucleotide Repeat Score (NRS) lower than 50,000, wherein said Nucleotide Repeat Score (NRS) is determined according to the formula:

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{l=1}^{k(n)} f_l(n)}}{N_{tot}},$$

wherein

N_{tot} is the length of said nucleotide sequence,

n is the length of a repeat within said nucleotide sequence, and

$f_i(n)$ is the frequency of said repeat of length n ,

wherein, if there is more than one repeat of length n , $k(n)$ is the number of said different sequences of said repeat of length n , otherwise $k(n)$ is 1 for said repeat of length n .

2. The nucleic acid molecule of item 1, wherein said encoded polypeptide consists of proline and alanine.
3. The nucleic acid molecule of item 2, wherein said proline residues constitute more than about 10 % and less than about 75 % of said encoded polypeptide.

- 3 -

4. The nucleic acid molecule of item 1, wherein said encoded polypeptide consists of proline, alanine and serine.
5. The nucleic acid molecule of item 4, wherein said proline residues constitute more than 4 % and less than 40 % of said encoded polypeptide.
6. The nucleic acid molecule of any one of items 1 to 5, wherein said Nucleotide Repeat Score (NRS) is lower than 100.
7. The nucleic acid molecule of any one of items 1 to 6, wherein said nucleic acid molecule has an enhanced genetic stability.
8. The nucleic acid molecule of any one of items 1 to 7, wherein said nucleotide sequence comprises said repeats, wherein said repeats have a maximum length $n_{\max}$, wherein $n_{\max}$ is determined according to the formula:

$$n_{\max} \leq 17 + \frac{N_{\text{tot}}}{600}$$

and wherein N_{tot} is the length of said nucleotide sequence.

9. The nucleic acid molecule of any one of items 1 to 8, wherein said repeats have a maximum length of about 14, 15, 16, or 17 nucleotides to about 55 nucleotides.
10. The nucleic acid molecule of any one of items 1 to 9, wherein said repeats have a maximum length corresponding to 50 % of the length of said nucleotide sequence.
11. The nucleic acid molecule of any one of items 1 to 10, wherein said encoded polypeptide comprises a repetitive amino acid sequence with a plurality of amino acid repeats, wherein no more than 9 consecutive amino acid residues are identical and wherein said polypeptide forms a random coil.
12. The nucleic acid molecule of any one of items 1 to 11, wherein said nucleic acid molecule is selected from the group consisting of:
 - (a) the nucleic acid molecule comprising at least one nucleotide sequence selected from the group consisting of SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21,

- 4 -

- SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 26 and SEQ ID NO: 27;
- (b) the nucleic acid molecule comprising at least one nucleotide sequence selected from the group consisting of SEQ ID NO: 28, SEQ ID NO: 29, SEQ ID NO: 30, SEQ ID NO: 31, SEQ ID NO: 32, SEQ ID NO: 33, SEQ ID NO: 34, SEQ ID NO: 35, SEQ ID NO: 36 and SEQ ID NO: 37.
 - (c) the nucleic acid molecule comprising the nucleotide sequence selected from the group consisting of SEQ ID NO: 38, SEQ ID NO: 39, SEQ ID NO: 40 and SEQ ID NO: 41;
 - (d) the nucleic acid molecule comprising the nucleotide sequence consisting of SEQ ID NO: 42, SEQ ID NO: 43, SEQ ID NO: 44, and/or SEQ ID NO: 45;
 - (e) the nucleic acid molecule hybridizing under stringent conditions to the complementary strand of the nucleotide sequence as defined in (a) or (b);
 - (f) the nucleic acid molecule comprising the nucleotide sequence having at least 56% identity to the nucleotide sequence as defined in any one of (a), (c) and (e);
 - (g) the nucleic acid molecule comprising the nucleotide sequence having at least 66.7% identity to the nucleotide sequence as defined in any one of (b), (d) and (e); and
 - (h) the nucleic acid molecule being degenerate as a result of the genetic code to the nucleotide sequence as defined in any one of (a) to (d).
13. The nucleic acid molecule of any one of items 1 to 12, wherein said nucleic acid molecule comprises two complementary 5'-overhangs, wherein the 5'-overhang on the coding strand is 5'-GCC, and wherein the 5'-overhang on the non-coding strand is 5'-GGC.
14. The nucleic acid molecule of any one of items 1 to 13 operably linked in the same reading frame to a nucleic acid encoding a biologically active protein.
15. The nucleic acid molecule of item 14, wherein said biologically active protein is a therapeutically effective protein.
16. The nucleic acid molecule of item 14 or 15, wherein said biologically active protein is selected from the group consisting of a binding protein, an antibody fragment, a cytokine, a growth factor, a hormone, an enzyme, a protein vaccine, a peptide vaccine, a peptide which consists of up to 50 amino acid residues or a peptidomimetic.

- 5 -

17. The nucleic acid molecule of item 16, wherein said binding protein is selected from the group consisting of antibodies, Fab fragments, Fab' fragments, F(ab')₂ fragments, single chain variable fragments (scFv), (single) domain antibodies, isolated variable regions of antibodies (VL and/or VH regions), CDRs, immunoglobulin domains, CDR-derived peptidomimetics, lectins, protein scaffolds, fibronectin domains, tenascin domains, protein A domains, SH3 domains, ankyrin repeat domains, and lipocalins.

18. The nucleic acid molecule of any one of items 14 to 17, wherein said biologically active protein is selected from the group consisting of interleukin 1 receptor antagonist, leptin, acid sphingomyelinase, adenosine deaminase, agalsidase alfa, alpha-1 antitrypsin, alpha atrial natriuretic peptide, alpha-galactosidase, alpha-glucosidase, alpha-N-acetylglucosaminidase, alteplase, amediase, amylin, amylin analog, anti-HIV peptide fusion inhibitor, arginine deiminase, asparaginase, B domain deleted factor VIII, bone morphogenetic protein, bradykinin antagonist, B-type natriuretic peptide, bouganin, growth hormone, chorionic gonadotropin, CD3 receptor antagonist, CD19 antagonist, CD20 antagonist, CD40 antagonist, CD40L antagonist, cerebroside sulfatase, coagulation factor VIIa, coagulation factor XIII, coagulation factor IX, coagulation factor X, complement component C3 inhibitor, complement component 5a antagonist, C-peptide, CTLA-4 antagonist, C-type natriuretic peptide, defensin, deoxyribonuclease I, EGFR receptor antagonist, epidermal growth factor, erythropoietin, exendin-4, ezrin peptide 1, FcγIIB receptor antagonist, fibroblast growth factor 21, follicle-stimulating hormone, gastric inhibitory polypeptide (GIP), GIP analog, glucagon, glucagon receptor agonist, glucagon-like peptide 1 (GLP-1), GLP-1 analog, glucagon-like peptide 2 (GLP-2), GLP-2 analog, gonadorelin, gonadotropin-releasing hormone agonist, gonadotropin-releasing hormone antagonist, gp120, gp160, granulocyte colony stimulating factor (G-CSF), granulocyte macrophage colony stimulating factor (GM-CSF), ghrelin, ghrelin analog, growth hormone, growth hormone-releasing hormone, hematide, hepatocyte growth factor, hepatocyte growth factor receptor (HGFR) antagonist, hepcidin antagonist, hepcidin mimetic, Her2/neu receptor antagonist, histrelin, hirudin, hsp70 antagonist, humanin, hyaluronidase, hydrolytic lysosomal glucocerebroside-specific enzyme, iduronate-2-sulfatase, IgE antagonists, insulin, insulin analog, insulin-like growth factor 1, insulin-like growth factor 2, interferon-alpha, interferon-alpha antagonist, interferon-alpha superagonist, interferon-alpha-n3, interferon-beta, interferon-gamma, interferon-lambda, interferon tau, interleukin, interleukin 2 fusion protein, interleukin-22 receptor subunit alpha (IL-22ra) antagonist, irisin, islet neogenesis associated protein, keratinocyte growth factor, Kv1.3 ion channel antagonists, lanthipeptide, lipase, luteinizing hormone, lutropin alpha, lysostaphin, mannosidase, N-acetylgalactosamine-6-sulfatase, N-

- 6 -

acetylglucosaminidase, neutrophil gelatinase-associated lipocalin, octreotide, ω -conotoxin, *Ornithodoros moubata* complement inhibitor, osteogenic protein-1, osteoprotegerin, oxalate decarboxylase, P128, parathyroid hormone, Phylomer, PD-1 antagonist, PDGF antagonist, phenylalanine ammonia lyase, platelet derived growth factor, proinsulin, protein C, relaxin, relaxin analog, secretin, RGD peptide, ribonuclease, senrebotase, serine protease inhibitor, soluble complement receptor type 1, soluble DCC receptor, soluble TACI receptor, soluble tumor necrosis factor I receptor (sTNF-RI), soluble tumor necrosis factor II receptor (sTNF-RII), soluble VEGF receptor Flt-1, soluble Fc γ IIB receptor, somatostatin, somatostatin analog, streptokinase, T-cell receptor ligand, tenecteplase, teriparatide, thrombomodulin alpha, thymosin alpha 1, toll like receptor inhibitor, tumor necrosis factor (TNF α), tumor necrosis factor α antagonist, uricase, vasoactive intestinal peptide, vasopressin, vasopressin analog, VEGF antagonist, von Willebrand factor.

19. A vector comprising the nucleic acid molecule of any one of items 1 to 18.
20. The vector of item 19, wherein said vector comprises an upstream recognition sequence that is recognized by an endonuclease restriction enzyme and a downstream recognition sequence that is recognized by an endonuclease restriction enzyme, and wherein said upstream recognition sequence and said downstream recognition sequence are in a reverse complementary orientation.
21. The vector of item 20, wherein said endonuclease restriction enzyme recognizing said downstream recognition sequence is different from the endonuclease restriction enzyme recognizing the upstream recognition sequence.
22. The vector of item 20 or 21, wherein said upstream recognition sequence comprises two recognition sequences for two different restriction enzymes.
23. The vector of any one of items 20 to 22, wherein said downstream recognition sequence is comprised in the upstream recognition sequence.
24. The vector of any one of items 20 to 23, wherein said upstream recognition sequence and/or said downstream recognition sequence are recognition sequences for restriction enzymes producing nucleotide overhangs.

- 7 -

25. The vector of any one of items 20 to 24, wherein said upstream recognition sequence and/or said downstream recognition sequence are recognition sites for restriction enzymes that cleave outside of the recognition sequence.
26. The vector of any one of items 20 to 25, wherein said restriction enzyme is a type IIS restriction enzyme.
27. The vector of any one of items 20 to 26, wherein said upstream recognition sequence has the nucleotide sequence "5'-GCTCTTC-3'" and/or wherein said downstream recognition sequence has the nucleotide sequence "5'-CTCTTC-3'".
28. The vector of any one of items 20 to 27, wherein said upstream recognition sequence is recognized by *SapI* and *EarI* and/or wherein said downstream recognition sequence is recognized by *EarI*.
29. The vector of any one of items 20 to 28, wherein said vector has a sequence given in SEQ ID NO: 48 or SEQ ID NO: 55.
30. A host comprising the nucleic acid molecule of any one of items 1 to 18 or transformed with the vector of any one of items 19 to 29.
31. The host of item 30, wherein the host is selected from the group consisting of a bacterium, a mammalian cell, an insect cell, an algal cell, a ciliate, a yeast and a plant cell.
32. The host of item 30 or 31, wherein the bacterium belongs to the genus *Escherichia*, *Corynebacterium*, *Pseudomonas* or *Bacillus*.
33. The host of item 32, wherein said bacterium is *Escherichia coli*, *Corynebacterium glutamicum*, *Pseudomonas fluorescens* or *Bacillus megaterium*.
34. The host of item 31, wherein said mammalian cell is a hamster cell.
35. The host of item 34, wherein said mammalian cell is a CHO cell.
36. The host of item 31, wherein said yeast belongs to the genus *Saccharomyces*, *Pichia*, *Hansenula* or *Kluyveromyces*.

- 8 -

37. The host of item 36, wherein said yeast is *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Pichia pastoris*, *Pichia methanolica*, *Hansenula polymorpha*, or *Kluyveromyces lactis*.
38. A method for preparing said nucleic acid molecule of any one of items 1 to 18, or said vector of any one of items 19 to 29, wherein the method comprises culturing/raising the host of any one of items 30 to 37 and optionally isolating the produced nucleic acid molecule and/or vector.
39. A method for preparing a polypeptide encoded by the nucleic acid molecule of any one of items 1 to 18, wherein the method comprises culturing/raising the host of any one of items 30 to 37 comprising the nucleic acid molecule of any one of items 1 to 18 and optionally isolating the produced polypeptide.
40. A method for preparing a drug conjugate, wherein said drug conjugate comprises the polypeptide encoded by the nucleic acid molecule of any one of items 1 to 18 and further comprises (i) a biologically active protein and/or (ii) a small molecule and/or (iii) a carbohydrate.
41. The method for preparing the drug conjugate of item 40, wherein said nucleic acid molecule has an enhanced genetic stability.
42. The method for preparing the drug conjugate of item 40 or 41, wherein said biologically active protein is a therapeutically effective protein.
43. The method for preparing the drug conjugate of any one of items 40 to 42, wherein said polypeptide mediates an enhanced in vivo and/or in vitro stability of said drug conjugate.
44. The method for preparing the drug conjugate of any one of items 40 to 43, wherein said biologically active protein is selected from the group consisting of a binding protein, an antibody fragment, a cytokine, a growth factor, a hormone, an enzyme, a protein vaccine, a peptide vaccine, a peptide which consists of up to 50 amino acid residues or a peptidomimetic.
45. The method for preparing the drug conjugate of any one of items 40 to 44, wherein said binding protein is selected from the group consisting of antibodies, Fab fragments, Fab' fragments, F(ab')₂ fragments, single chain variable fragments (scFv), (single) domain

- 9 -

antibodies, isolated variable regions of antibodies (VL and/or VH regions), CDRs, immunoglobulin domains, CDR-derived peptidomimetics, lectins, protein scaffolds, fibronectin domains, tenascin domains, protein A domains, SH3 domains, ankyrin repeat domains, and lipocalins.

46. The method for preparing the drug conjugate of any one of items 40 to 45, wherein said biologically active protein is selected from the group consisting of interleukin 1 receptor antagonist, leptin, acid sphingomyelinase, adenosine deaminase, agalsidase alfa, alpha-1 antitrypsin, alpha atrial natriuretic peptide, alpha-galactosidase, alpha-glucosidase, alpha-N-acetylglucosaminidase, alteplase, amediase, amylin, amylin analog, anti-HIV peptide fusion inhibitor, arginine deiminase, asparaginase, B domain deleted factor VIII, bone morphogenetic protein, bradykinin antagonist, B-type natriuretic peptide, bouganin, growth hormone, chorionic gonadotropin, CD3 receptor antagonist, CD19 antagonist, CD20 antagonist, CD40 antagonist, CD40L antagonist, cerebroside sulfatase, coagulation factor VIIa, coagulation factor XIII, coagulation factor IX, coagulation factor X, complement component C3 inhibitor, complement component 5a antagonist, C-peptide, CTLA-4 antagonist, C-type natriuretic peptide, defensin, deoxyribonuclease I, EGFR receptor antagonist, epidermal growth factor, erythropoietin, exendin-4, ezrin peptide 1, FcγIIb receptor antagonist, fibroblast growth factor 21, follicle-stimulating hormone, gastric inhibitory polypeptide (GIP), GIP analog, glucagon, glucagon receptor agonist, glucagon-like peptide 1 (GLP-1), GLP-1 analog, glucagon-like peptide 2 (GLP-2), GLP-2 analog, gonadorelin, gonadotropin-releasing hormone agonist, gonadotropin-releasing hormone antagonist, gp120, gp160, granulocyte colony stimulating factor (G-CSF), granulocyte macrophage colony stimulating factor (GM-CSF), grehlin, grehlin analog, growth hormone, growth hormone-releasing hormone, hematide, hepatocyte growth factor, hepatocyte growth factor receptor (HGFR) antagonist, hepcidin antagonist, hepcidin mimetic, Her2/neu receptor antagonist, histrelin, hirudin, hsp70 antagonist, humanin, hyaluronidase, hydrolytic lysosomal glucocerebroside-specific enzyme, iduronate-2-sulfatase, IgE antagonists, insulin, insulin analog, insulin-like growth factor 1, insulin-like growth factor 2, interferon-alpha, interferon-alpha antagonist, interferon-alpha superagonist, interferon-alpha-n3, interferon-beta, interferon-gamma, interferon-lambda, interferon tau, interleukin, interleukin 2 fusion protein, interleukin-22 receptor subunit alpha (IL-22ra) antagonist, irisin, islet neogenesis associated protein, keratinocyte growth factor, Kv1.3 ion channel antagonists, lanthipeptide, lipase, luteinizing hormone, lutropin alpha, lysostaphin, mannosidase, N-acetylgalactosamine-6-sulfatase, N-acetylglucosaminidase, neutrophil gelatinase-associated lipocalin, octreotide, ω-conotoxin, Ornithodoros moubata complement inhibitor, osteogenic protein-1,

- 10 -

osteoprotegerin, oxalate decarboxylase, P128, parathyroid hormone, Phylomer, PD-1 antagonist, PDGF antagonist, phenylalanine ammonia lyase, platelet derived growth factor, proinsulin, protein C, relaxin, relaxin analog, secretin, RGD peptide, ribonuclease, senrebotase, serine protease inhibitor, soluble complement receptor type 1, soluble DCC receptor, soluble TACI receptor, soluble tumor necrosis factor I receptor (sTNF-RI), soluble tumor necrosis factor II receptor (sTNF-RII), soluble VEGF receptor Flt-1, soluble FcγIIB receptor, somatostatin, somatostatin analog, streptokinase, T-cell receptor ligand, tenecteplase, teriparatide, thrombomodulin alpha, thymosin alpha 1, toll like receptor inhibitor, tumor necrosis factor (TNFα), tumor necrosis factor α antagonist, uricase, vasoactive intestinal peptide, vasopressin, vasopressin analog, VEGF antagonist, von Willebrand factor.

47. The method for preparing the drug conjugate of any one of items 40 to 46, wherein said small molecule is selected from the group consisting of angiogenesis inhibitors, anti-allergic drugs, anti-emetic drugs, anti-depressant drugs, anti-hypertensive drugs, anti-inflammatory drugs, anti-infective drugs, anti-psychotic drugs, anti-proliferative (cytotoxic and cytostatic) drugs, calcium antagonists and other circulatory organ drugs, cholinergic agonists, drugs acting on the central nervous system, drugs acting on the respiratory system, hormones, steroids, polyketides, carbohydrates, oligosaccharides, nucleic acids, nucleic acid derivatives, antisense nucleic acids, small interference RNAs (siRNAs), micro RNA (miR) inhibitors, microRNA mimetics, DNA aptamers and RNA aptamers.
48. A polypeptide as obtained or obtainable by the method of item 39, a drug conjugate as obtained by the method of any one of items 40 to 47.
49. Method for sequencing of the nucleic acid molecule of any one of items 1 to 18.
50. Method for amplification of the nucleic acid molecule of any one of items 1 to 18.
51. Method for cloning of the nucleic acid molecule of any one of items 1 to 18.
52. A method for selecting a genetically stable nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein said nucleotide sequence has a length of at least 300 nucleotides,
the method comprising a step of selecting a nucleic acid molecule comprising a nucleotide sequence having a Nucleotide Repeat Score (NRS) lower than 50,000,

- 11 -

wherein said Nucleotide Repeat Score (NRS) is determined according to the formula:

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{i=1}^{k(n)} f_i(n)}}{N_{tot}},$$

wherein

N_{tot} is the length of said nucleotide sequence,

n is the length of a repeat within said nucleotide sequence, and

$f_i(n)$ is the frequency of said repeat of length n ,

wherein if there is more than one repeat of length n , $k(n)$ is the number of said different sequences of said repeat of length n , otherwise $k(n)$ is 1 for said repeat of length n .

In certain aspects the present application relates to the following items:

1. A nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein the nucleotide sequence of said nucleic acid has a length of at least 300 nucleotides, wherein said nucleotide sequence has a Nucleotide Repeat Score (NRS) lower than 50,000, wherein said Nucleotide Repeat Score (NRS) is determined according to the formula:

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{i=1}^{k(n)} f_i(n)}}{N_{tot}},$$

wherein

N_{tot} is the length of said nucleotide sequence,

n is the length of a repeat within said nucleotide sequence, and

$f_i(n)$ is the frequency of said repeat of length n ,

wherein if there is more than one repeat of length n , $k(n)$ is the number of said different sequences of said repeat of length n , otherwise $k(n)$ is 1 for said repeat of length n .

- 12 -

2. The nucleic acid molecule of item 1, wherein said encoded polypeptide consists of proline and alanine, wherein said proline residues constitute more than about 10 % and less than about 75 % of said encoded polypeptide.
3. The nucleic acid molecule of item 1, wherein said encoded polypeptide consists of proline, alanine and serine, wherein said proline residues constitute more than 4% and less than 40% of said encoded polypeptide.
4. The nucleic acid molecule of any one of items 1 to 3, wherein said Nucleotide Repeat Score (NRS) is lower than 100.
5. The nucleic acid molecule of any one of items 1 to 4, wherein said nucleic acid molecule has an enhanced genetic stability.
6. The nucleic acid molecule of any one of items 1 to 5, wherein said nucleotide sequence comprises said repeats, wherein said repeats have a maximum length $n_{\max}$, wherein $n_{\max}$ is determined according to the formula:

$$n_{\max} \leq 17 + \frac{N_{\text{tot}}}{600}$$

and wherein N_{tot} is the length of said nucleotide sequence.

7. The nucleic acid molecule of any one of items 1 to 6, wherein said repeats have a maximum length of about 14, 15, 16, or 17 nucleotides to about 55 nucleotides.
8. The nucleic acid molecule of any one of items 1 to 7, wherein said encoded polypeptide comprises a repetitive amino acid sequence with a plurality of amino acid repeats, wherein no more than 9 consecutive amino acid residues are identical and wherein said polypeptide forms a random coil.

- 13 -

9. The nucleic acid molecule of any one of items 1 to 8, wherein said nucleic acid molecule is selected from the group consisting of:
 - (a) the nucleic acid molecule comprising at least one nucleotide sequence selected from the group consisting of SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 26 and SEQ ID NO: 27;
 - (b) the nucleic acid molecule comprising at least one nucleotide sequence selected from the group consisting of SEQ ID NO: 28, , SEQ ID NO: 29, SEQ ID NO: 30, SEQ ID NO: 31, SEQ ID NO: 32, SEQ ID NO: 33, SEQ ID NO: 34, SEQ ID NO: 35, SEQ ID NO: 36 and SEQ ID NO: 37;
 - (c) the nucleic acid molecule comprising the nucleotide sequence selected from the group consisting of SEQ ID NO: 38, SEQ ID NO: 39, SEQ ID NO: 40 and SEQ ID NO: 41;
 - (d) the nucleic acid molecule comprising the nucleotide sequence consisting of SEQ ID NO: 42, SEQ ID NO: 43 SEQ ID NO: 44 and/or SEQ ID NO: 45
 - (e) the nucleic acid molecule hybridizing under stringent conditions to the complementary strand of the nucleotide sequence as defined in (a) or (b);
 - (f) the nucleic acid molecule comprising the nucleotide sequence having at least 56% identity to the nucleotide sequence as defined in any one of (a), (c) and (e);
 - (g) the nucleic acid molecule comprising the nucleotide sequence having at least 66.7% identity to the nucleotide sequence as defined in any one of (b), (d) and (e); and
 - (h) the nucleic acid molecule being degenerate as a result of the genetic code to the nucleotide sequence as defined in any one of (a) to (d).
10. The nucleic acid molecule of any one of items 1 to 9 operably linked in the same reading frame to a nucleic acid encoding a biologically active protein.
11. The nucleic acid molecule of item 10,
wherein said biologically active protein is selected from the group consisting of a binding protein, an antibody fragment, a cytokine, a growth factor, a hormone, an enzyme, a protein vaccine, a peptide vaccine, a peptide which consists of up to 50

- 14 -

amino acid residues or a peptidomimetic, wherein said binding protein is selected from the group consisting of antibodies, Fab fragments, Fab' fragments, F(ab')₂ fragments, single chain variable fragments (scFv), (single) domain antibodies, isolated variable regions of antibodies (VL and/or VH regions), CDRs, immunoglobulin domains, CDR-derived peptidomimetics, lectins, protein scaffolds, fibronectin domains, tenascin domains, protein A domains, SH3 domains, ankyrin repeat domains, and lipocalins.

12. The nucleic acid molecule of item 10 or 11, wherein said biologically active protein is selected from the group consisting of acid sphingomyelinase, adenosine deaminase, agalsidase alfa, alpha-1 antitrypsin, alpha atrial natriuretic peptide, alpha-galactosidase, alpha-glucosidase, alpha-N-acetylglucosaminidase, alteplase, amediase, amylase, amylase analog, anti-HIV peptide fusion inhibitor, arginine deiminase, asparaginase, B domain deleted factor VIII, bone morphogenetic protein, bradykinin antagonist, B-type natriuretic peptide, bouganin, growth hormone, chorionic gonadotropin, CD3 receptor antagonist, CD19 antagonist, CD20 antagonist, CD40 antagonist, CD40L antagonist, cerebroside sulfatase, coagulation factor VIIa, coagulation factor XIII, coagulation factor IX, coagulation factor X, complement component C3 inhibitor, complement component 5a antagonist, C-peptide, CTLA-4 antagonist, C-type natriuretic peptide, defensin, deoxyribonuclease I, EGFR receptor antagonist, epidermal growth factor, erythropoietin, exendin-4, ezrin peptide 1, FcγIIb receptor antagonist, fibroblast growth factor 21, follicle-stimulating hormone, gastric inhibitory polypeptide (GIP), GIP analog, glucagon, glucagon receptor agonist, glucagon-like peptide 1 (GLP-1), GLP-1 analog, glucagon-like peptide 2 (GLP-2), GLP-2 analog, gonadorelin, gonadotropin-releasing hormone agonist, gonadotropin-releasing hormone antagonist, gp120, gp160, granulocyte colony stimulating factor (G-CSF), granulocyte macrophage colony stimulating factor (GM-CSF), ghrelin, ghrelin analog, growth hormone, growth hormone-releasing hormone, hematide, hepatocyte growth factor, hepatocyte growth factor receptor (HGFR) antagonist, hepcidin antagonist, hepcidin mimetic, Her2/neu receptor antagonist, histrelin, hirudin, hsp70 antagonist, humanin, hyaluronidase, hydrolytic lysosomal glucocerebroside-specific enzyme, iduronate-2-sulfatase, IgE antagonists, insulin, insulin analog, insulin-like growth factor 1, insulin-like growth factor 2, interferon-alpha, interferon-alpha antagonist, interferon-alpha superagonist,

- 15 -

interferon-alpha-n3, interferon-beta, interferon-gamma, interferon-lambda, interferon tau, interleukin, interleukin 1 receptor antagonist, interleukin 2 fusion protein, interleukin-22 receptor subunit alpha (IL-22ra) antagonist, irisin, islet neogenesis associated protein, keratinocyte growth factor, Kv1.3 ion channel antagonists, lanthipeptide, leptin, lipase, luteinizing hormone, lutropin alpha, lysostaphin, mannosidase, N-acetylgalactosamine-6-sulfatase, N-acetylglucosaminidase, neutrophil gelatinase-associated lipocalin, octreotide, ω -conotoxin, Ornithodoros moubata complement inhibitor, osteogenic protein-1, osteoprotegerin, oxalate decarboxylase, P128, parathyroid hormone, Phylomer, PD-1 antagonist, PDGF antagonist, phenylalanine ammonia lyase, platelet derived growth factor, proinsulin, protein C, relaxin, relaxin analog, secretin, RGD peptide, ribonuclease, senrebotase, serine protease inhibitor, soluble complement receptor type 1, soluble DCC receptor, soluble TACI receptor, soluble tumor necrosis factor I receptor (sTNF-RI), soluble tumor necrosis factor II receptor (sTNF-RII), soluble VEGF receptor Flt-1, soluble Fc γ IIb receptor, somatostatin, somatostatin analog, streptokinase, T-cell receptor ligand, tenecteplase, teriparatide, thrombomodulin alpha, thymosin alpha 1, toll like receptor inhibitor, tumor necrosis factor (TNF α), tumor necrosis factor α antagonist, uricase, vasoactive intestinal peptide, vasopressin, vasopressin analog, VEGF antagonist, von Willebrand factor.

13. A host comprising the nucleic acid molecule of any one of items 1 to 12 or transformed with a vector comprising the nucleic acid molecule of any one of items 1 to 12.
14. A method for preparing a polypeptide encoded by the nucleic acid molecule of any one of items 1 to 12, wherein the method comprises culturing/raising the host of item 13 and optionally isolating the produced polypeptide.
15. A method for preparing a drug conjugate, wherein said drug conjugate comprises the polypeptide encoded by the nucleic acid molecule of any one of items 1 to 12 and further comprises (i) a biologically active protein and/or (ii) a small molecule and/or (iii) a carbohydrate.

- 16 -

The present invention relates to nucleic acid molecules with low repetitive nucleotide sequences encoding polypeptides consisting of proline, alanine and, optionally, serine. Such polypeptides are also termed herein PA-rich or proline/alanine-rich polypeptides. The inventive nucleic acid molecules provided herein have few or no nucleotide repeats of a certain maximum length, such as a maximum length of about 14, 15, 16 or 17 nucleotides per nucleotide repeat sequence. Furthermore, the PA-rich coding nucleotide sequence has an overall length of at least 300 nucleotides and the individual nucleotide repeats within this coding sequence have an individual maximal length of 14, 15, 16, 17, about 20, about 25, about 30, about 35, about 40, about 45, about 50 or about 55 nucleotides.

In one aspect, the nucleic acid molecule of the invention comprises a nucleotide sequence encoding a repetitive amino acid sequence (e.g. a PA-rich polypeptide), wherein said nucleotide sequence of said nucleic acid molecule has a length of at least 300 nucleotides, wherein said nucleotide sequence has a Nucleotide Repeat Score (NRS) lower than 50,000, and wherein said Nucleotide Repeat Score (NRS) is determined according to the formula:

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{i=1}^{k(n)} f_i(n)}}{N_{tot}}.$$

In this formula, N_{tot} is the length of said nucleotide sequence, n is the length of a repeat within said nucleotide sequence, and $f_i(n)$ is the frequency of said repeat of length n , and wherein, if there is more than one repeat of length n , $k(n)$ is the number of said different sequences of said repeat of length n , otherwise $k(n)$ is 1 for said repeat of length n . A definition follows below.

In a certain aspect, the nucleic acid molecule of the invention encodes a polypeptide consisting of proline, alanine and, optionally, serine, wherein a nucleotide sequence stretch which repeatedly occurs within the nucleotide sequence of said nucleic acid molecule (i.e. a “repeat”) has maximally a length of 14, 15, 16, 17, about 20, about 25, about 30, about 35, about 40, about 45, about 50 or about 55 nucleotides. In other words, the nucleic acid molecule comprises a nucleotide sequence encoding a PA-rich polypeptide, wherein said coding nucleotide sequence comprises nucleotide repeats having a maximum length of 14, 15, 16, 17, about 20, about 25, about 30, about 35, about 40, about 45, about 50 or about 55 nucleotides. The inventive nucleic acid molecules/sequences can also comprise additional coding sequences, like inter alia biologically or pharmacologically active proteins.

It is demonstrated in the appended examples that the low repetitive nucleic acid molecules of the invention are advantageous compared to highly repetitive nucleic acid molecules of the

- 17 -

prior art. In particular, the genetic stability of the low repetitive nucleic acid molecules provided herein is improved as documented herein and shown in the appended examples. In order to assess the genetic stability of prior art repetitive nucleic acid molecules encoding PA-rich polypeptides, a vector was constructed comprising a prior art repetitive nucleic acid molecule encoding a repetitive proline/alanine-rich sequence (PAS#1a(600); SEQ ID NO: 12; Example 6, Fig. 2B) which is composed of multiple 60mer nucleotide sequence units as disclosed in WO 2008/155134. The vector is termed herein "pASK75-PAS#1a(600)-IL1Ra" (SEQ ID NO: 51). The host (*E. coli*) was transformed with the vector and cultivated for several days, e.g., 7 days. On day 7 after a continuous growth over approximately 70 generations, cells were plated on LB/Amp agar, clones were picked, and plasmid preparations were performed. Plasmids were analysed using restriction enzymes and subsequent agarose gel electrophoresis (Fig. 5). Four of five analyzed clones of pASK75-PAS#1a(600)-IL1Ra showed shortened nucleic acid fragments encoding the proline/alanine-rich sequences (Fig. 5 lanes 1-5).

Accordingly, prior art repetitive nucleic acid molecules encoding repetitive proline/alanine-rich sequences are genetically unstable. In other words, prior art nucleic acid molecules have a low in vivo stability. Without being bound by theory the genetic instability might be a result of homologous recombination. Due to the genetic instability and shortening of the gene cassette, the resulting proline/alanine-rich amino acid sequence will be altered, too. Thus, the proline/alanine-rich amino acid sequence encoded by the unstable plasmids during long-term cultivation will be different from that encoded by the original plasmid. Therefore, there is considerable risk that the polypeptides obtained during long-term cultivation using prior art repetitive nucleic acid molecules will not be the desired polypeptides.

Moreover, there is a risk that the resulting polypeptide composition may comprise a variety of different polypeptides (e.g. proline/alanine-rich polypeptides of various sizes, lengths and/or sequences) which is contrary to the desired conformity of biological products, particularly for biopharmaceutical use. Thus, genetic instability of repetitive prior art nucleic acid molecules encoding proline/alanine-rich polypeptides can lead to decreased quality of the biological end-product, rendering the production thereof inconvenient and unreliable, especially under regulatory aspects for therapeutic application.

In the present invention, the problem of genetic instability is solved by designing nucleic acid molecules with low internal nucleotide repeats. However, given the low number of nucleotide triplet codons available to encode proline/alanine-rich amino acid sequences, this was not a trivial task.

As shown in the appended non-limiting examples, the nucleic acid molecules of the invention

- 18 -

avoid the above recited disadvantages; see e.g. Examples 5 and 6 and Fig. 4 and Fig. 5 lane 6-10. Nevertheless, like the prior art nucleic acid molecules, the nucleic acid molecules of the invention encode proline/alanine-rich polypeptides containing a plurality of proline/alanine-rich repeats. Yet, in strong contrast to the prior art nucleic acid molecules, the nucleic acid molecules of the present invention have a low repetitive nucleotide sequence (i.e. they contain few and/or only short nucleotide repeats).

Illustrative Example 5 shows the preparation of an exemplary vector of the invention comprising a low repetitive nucleotide sequence encoding a proline/alanine-rich polypeptide. The low repetitive nucleotide sequence termed "PAS#1f/1c/1b(600)" as used in the exemplary vector is shown in SEQ ID NO: 38. The resulting plasmid was designated "pASK75-PAS#1f/1c/1b(600)-IL1Ra" (SEQ ID NO: 50) and is shown in Fig. 4.

The plasmid of the invention, "pASK75-PAS#1f/1c/1b(600)-IL1Ra" was subjected to the same cultivation as described above in relation to the vector "pASK75-PAS#1a(600)-IL1Ra", the latter vector comprising a prior art repetitive nucleic acid molecule encoding a repetitive proline/alanine-rich sequence (PAS#1a(600); SEQ ID NO: 12), which is composed of multiple 60mer nucleotide sequence units as disclosed in WO 2008/155134 (Figure 1A). In contrast to "pASK75-PAS#1a(600)-IL1Ra", the plasmid of the invention, "pASK75-PAS#1f/1c/1b(600)-IL1Ra", showed high genetic stability: all analysed clones of "pASK75-PAS#1f/1c/1b(600)" only showed the expected bands at 3093 bp and 2377 bp (Fig. 5 lanes 6-10), indicating a high genetic stability of the low repetitive PAS#1f/1c/1b(600) gene cassette comprising 1800 base pairs and encoding the proline/alanine-rich sequence PAS#1. Due to this high genetic stability, the disadvantages of prior art nucleic acid molecules are avoided. This clearly demonstrates that the nucleic acid molecules of the invention comprising a low repetitive nucleotide sequence are useful for the convenient and reliable biosynthesis of proline/alanine-rich polypeptides and/or corresponding fusion proteins.

Illustrative Example 4 demonstrates a further advantage of the nucleic acid molecules provided in accordance with the present invention. Here, the exemplary nucleic acid molecule comprising a low repetitive nucleotide sequence of the invention (termed PAS#1f/1c/1b(600) cassette; SEQ ID NO: 38; see e.g. Example 1) was subjected to automated DNA sequencing. As result, a clearly defined and error-free electropherogram comprising more than 900 base pairs (Fig. 3) was obtained, which showed no signs of unspecific primer binding. Thus, in contrast to repetitive nucleotide sequences, which can only be reliably sequenced with primers that hybridize upstream or downstream of a corresponding cloned gene cassette, long low repetitive DNA fragments encoding proline/alanine-rich sequences can be fully sequenced in a facile way. In this case, internally binding primers can also be applied, thus generating multiple

- 19 -

overlapping sequence reads if desired; notably, the use of such internally hybridizing primers does not result in unique sequencing patterns in the case of repetitive nucleotide sequences. Accordingly, the nucleic acid molecules of the invention comprising the low repetitive nucleotide sequence circumvent these sequencing problems.

In summary, the present invention has, *inter alia*, the following advantages over prior art nucleic acid molecules that comprise highly repetitive sequences. The advantageous low repetitive nucleotide sequences of the present invention can be fully sequenced without further ado in contrast to nucleic acid molecules of the prior art. A further advantage of the nucleic acid molecules of the present invention is that they have improved amplification properties, e.g. via the polymerase chain reaction (PCR), due to the low repetitiveness. Furthermore, the nucleic acid molecules of the invention improve the cloning procedure compared to nucleotide sequences comprising repeated/repetitive sequences. A particular advantage of the herein provided nucleic acid molecules is that they have an improved genetic stability compared to the highly repetitive nucleic acid molecules of the prior art. This allows a reliable production of proline/alanine-rich polypeptides and/or fusion proteins thereof.

The characterizing feature of the nucleic acid molecules of the present invention is that the nucleotide sequences encoding a polypeptide consisting of proline, alanine and, optionally, serine are “low repetitive nucleotide sequences”, which confers the advantageous technical effects described above. In the appended Examples, methods are demonstrated that can be employed to analyze whether a nucleic acid molecule comprises a low repetitive nucleotide sequence according to the invention. In particular, the appended Examples provide a score termed herein “Nucleotide Repeat Score (NRS)”. This Nucleotide Repeat Score (NRS) is herein determined according to the formula as discussed above:

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{l=1}^{k(n)} f_l(n)}}{N_{tot}},$$

wherein N_{tot} is the length of said nucleotide sequence, n is the length of a repeat within said nucleotide sequence, and $f_l(n)$ is the frequency of said repeat of length n , and wherein, if there is more than one repeat of length n , $k(n)$ is the number of said different sequences of said repeat of length n , otherwise $k(n)$ is 1 for said repeat of length n . This score is described in detail below and is illustrated in the appended examples.

The NRS allows the skilled person to select low repetitive nucleotide sequences for use in the present invention. In other words, the NRS provides a tool for determining the degree of

- 20 -

repetitiveness of a nucleotide sequence. In order to automatically identify repeats and to calculate the NRS, the algorithm termed NRS-Calculator provided herein can be employed.

As demonstrated in the appended illustrative Examples, e.g., Example 13, several prior art nucleic acids encoding proline/alanine-rich sequences were compared to low repetitive nucleic acids encoding proline/alanine-rich sequences according to this invention using the NRS-Calculator described in Example 14. For example, the NRS of the following prior art sequences was determined: PAS#1a(200) disclosed in WO 2008/155134, PA#1a(200) disclosed in WO2011144756, [(AP)₅]₂₀APA disclosed in US 20060252120, [AAPAPAPAP]₁₀AS published under GenBank accession number DQ399411a, the large tegument protein of the macacine herpesvirus 1 published under GenBank accession number NP_851896. In addition, histograms showing the frequencies of the nucleotide repeats plotted against the respective length were determined for these prior art nucleotide sequences and for the low repetitive nucleotide sequences of the present invention such as PAS#1b(200) (SEQ ID NO: 19) or PA#1e/1d/1c/1b(800) (SEQ ID NO: 44) (Fig. 9). The histograms of the prior art nucleotide sequences revealed a highly repetitive nature. In contrast, the histograms of the low repetitive nucleotide sequences of the invention, e.g., PAS#1b(200) and PA#1e/1d/1c/1b(800), demonstrate only a few repeats with a maximum length of 14 nucleotides; see e.g. Fig. 9F-I.

The difference in repetitiveness between the prior art nucleotide sequences and the inventive nucleotide sequences becomes even more evident when comparing their Nucleotide Repeat Scores. The prior art sequences analyzed herein reveal an NRS above 80000 (Table 2). In contrast, the exemplary inventive nucleotide sequences demonstrate low Nucleotide Repeat Scores, e.g., below 34; see Table 1. Consequently, it is clearly proven herein that the repeat quality of the nucleotide sequences encoding proline/alanine-rich sequences of the invention is much higher compared to prior art sequences, with both fewer and shorter nucleotide sequence repeats. Accordingly, the nucleic acid molecules of the present invention have low repetitive sequences.

As indicated above, the skilled person is aware of several alternatives to analyze the degree of repetitiveness of a nucleic acid molecule. It is shown in the appended examples that the repetitiveness of the nucleic acid molecules of the invention and of those in the prior art was also analyzed by dot plot analysis; see e.g. Example 3. The dot plot analysis was performed for nucleic acids encoding the repetitive proline/alanine-rich sequence PA#3a (SEQ ID NO: 15; Fig. 2A) disclosed in WO 2011/144756, PAS#1 (SEQ ID NO: 11; Fig. 2B) disclosed in WO 2008/155134, an encoded [(AP)₅]_n multimer (SEQ ID NO: 16) disclosed in US2006/0252120 and a repetitive proline/alanine-rich region of the very large tegument protein of *Macacine herpesvirus 1* gene, published under the GenBank Accession Number AAP41454.1 (SEQ ID

- 21 -

NO: 18; Fig. 2C). In the appended Examples, the "dottup" tool of the Geneious software package version 8.1 (Biomatters, Auckland, New Zealand) was employed. The algorithm aligns the respective sequence to itself and applies a repeat window of, e.g., 14 or 15 nucleotides. The dot plots of the prior art nucleotide sequences were compared to dot plots of the low repetitive nucleotide sequences of the invention, e.g., the units PA#3b(200) (SEQ ID NO: 36), PA#1b (SEQ ID NO: 28) or the assembled low repetitive nucleotide sequences PAS#1f/1c/1b(600) (SEQ ID NO: 38) and PAS#1d/1f/1c/1b(800) (SEQ ID NO: 39). Whereas all analyzed prior art sequences revealed a highly repetitive nature on the nucleotide sequence level as illustrated by black parallel diagonal lines (Fig. 2A-C), dot plots of the exemplary nucleotide sequences according to the invention show no or only a few scattered repeats of a maximal length of 14 nucleotides (black lines) within the entire nucleotide sequence of 600 nucleotides (Fig. 2A,C), 1800 nucleotides or 2400 nucleotides, respectively (Fig. 2B). Accordingly, it is proven herein that the nucleotide sequences provided herein are low repetitive nucleotide sequences comprising no or only few short repeats.

In summary, the appended examples clearly demonstrate that the nucleic acid molecules of the invention have low repetitive nucleotide sequences while encoding a proline/alanine-rich amino acid repeat sequence. It is further demonstrated herein that the "low repetitiveness" or "low repetitiveness" of the nucleic acid molecules provided herein can readily be assessed by alternative strategies, e.g., the Nucleotide Repeat Score provided herein or methods known to the skilled person such as dot plot analysis. Alternatively, a skilled person can identify nucleotide sequence repeats and, thus, the degree of repetitiveness of the nucleotide sequence either manually or with the aid of generic software programs such as the Visual Gene Developer (Jung (2011) BMC Bioinformatics 12:340), or the Repfind tool (Betley (2002) Curr Biol 12:1756-1761). Thereby, the nucleic acid molecules of the invention having unexpected advantageous properties can easily be distinguished from prior art nucleic acid molecules lacking these characteristics.

As described above, the low repetitive nucleotide sequence of the invention can either have an NRS of lower than 50,000 or it can have repeats with a maximum length of about 17, about 20, about 25, about 30, about 35, about 40, about 45, about 50 or about 55 nucleotides. The low repetitive nucleotide sequence of the invention has a length of at least 100 nucleotides, preferably of at least 150, more preferably of at least 200, even more preferably of at least 300 nucleotides, even more preferably of at least 350 nucleotides, even more preferably of at least 600 nucleotides, even more preferably of at least 900, even more preferably of at least 1200, even more preferably of at least 1500 nucleotides, or most preferably of at least 1800 nucleotides. In other words, the nucleic acid molecule of the invention comprises or consists of a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally,

- 22 -

serine, wherein the nucleotide sequence of said nucleic acid has a length of at least 100 nucleotides, preferably of at least 150, more preferably of at least 200, even more preferably of at least 300 nucleotides, even more preferably of at least 350 nucleotides, even more preferably of at least 400 or 500 nucleotides, even more preferably of at least 600 nucleotides, even more preferably of at least 700 or 800 nucleotides, even more preferably of at least 900 nucleotides, even more preferably of at least 1000 or 1100, even more preferably of at least 1200 nucleotides (e.g. 1203 nucleotides), even more preferably of at least 1300 or 1400 nucleotides, even more preferably of at least 1500 nucleotides, even more preferably of at least 1600 or 1700 nucleotides, or most preferably of at least 1800 nucleotides.

The nucleic acid molecule of the invention can comprise or consist of a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein the nucleotide sequence of said nucleic acid has a length of maximally 5000 nucleotides, preferably maximally 4800 nucleotides, 3600 nucleotides, or 2400 nucleotides. The nucleic acid molecule of the invention can comprise or consist of a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein the nucleotide sequence of said nucleic acid has a length of maximally 5000 nucleotides, 4900 nucleotides, 4800 nucleotides, 4700 nucleotides, 4600 nucleotides, 4500 nucleotides, 4400 nucleotides, 4300 nucleotides, 4200 nucleotides, 4100 nucleotides, 4000 nucleotides, 3900 nucleotides, 3800 nucleotides, 3700 nucleotides, 3600 nucleotides, 3500 nucleotides, 3400 nucleotides, 3300 nucleotides, 3200 nucleotides, 3100 nucleotides, 3000 nucleotides, 2900 nucleotides, 2800 nucleotides, 2700 nucleotides, 2600 nucleotides, 2500 nucleotides, 2400 nucleotides, 2300 nucleotides, 2200 nucleotides, 2100 nucleotides, 2000 nucleotides, or maximally 1900 nucleotides.

In a particularly preferred aspect, the nucleic acid molecule of the invention can comprise or consist of a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein the nucleotide sequence of said nucleic acid has a length of 1200 to 3600 nucleotides.

Furthermore, the low repetitive nucleotide sequence has a Nucleotide Repeat Score (NRS) that is lower than 50,000, preferably lower than 40,000, more preferably lower than 30,000, more preferably lower than 20,000, more preferably lower than 10,000, more preferably lower than 1000, more preferably lower than 500, even more preferably lower than 100. Particularly preferred are low repetitive nucleotide sequences having a Nucleotide Repeat Score (NRS) of lower than 50, more preferably lower than 48, more preferably lower than 45, more preferably lower than 43, more preferably lower than 40, more preferably lower than 38, or most preferably lower than 35. In other words, the nucleic acid molecule of the invention comprises a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally,

- 23 -

serine, wherein said nucleotide sequence has a Nucleotide Repeat Score (NRS) lower than 50,000, preferably lower than 40,000, more preferably lower than 30,000, more preferably lower than 20,000, more preferably lower than 10,000, more preferably lower than 1000, more preferably lower than 500, even more preferably lower than 400, 300, 200, and even more preferably lower than 100. Particularly preferred are nucleic acid molecules comprising a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein said nucleotide sequence has a Nucleotide Repeat Score (NRS) lower than 90, 80, 70, 60, more preferably lower than 50, more preferably lower than 48, more preferably an NRS score of 45 or lower than 45, more preferably lower than 43, more preferably lower than 40, more preferably an NRS score of 39, 38, 37, or 36 or lower than 39, 38, 37, or 36, or most preferably an NRS score of 35 or lower than 35. Even more particularly preferred are nucleic acid molecules comprising a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein said nucleotide sequence has a Nucleotide Repeat Score (NRS) of 34, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, or 8 or an NRS lower than 34, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, or 8.

As discussed above, the “Nucleotide Repeat Score “ or “NRS” can be determined according to the following formula:

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{i=1}^{k(n)} f_i(n)}}{N_{tot}}.$$

Accordingly, the NRS is defined as the sum, over all possible repeat lengths (from $n = 4$ up to $N_{tot} - 1$), of each squared repeat length (n) multiplied with the square root of its overall frequency, divided through the total length of the analyzed nucleotide sequence (N_{tot}). In other words, the NRS is the sum of the squared length of repeats within said nucleotide sequence multiplied with the square root of the sum of the frequency of said repeat of length n ($f_i(n)$), wherein, if there is more than one repeat of length n , $k(n)$ is the number of said different sequences of said repeat of length n , otherwise $k(n)$ is 1 for said repeat of length n ; and wherein said sum is divided by the total length of said nucleotide sequence.

As used herein, “ N_{tot} ” is the overall length of said nucleotide sequence encoding the polypeptide consisting of proline, alanine and, optionally, serine. The length N_{tot} is also the number of nucleotides of said nucleotide sequence. Thus, N_{tot} is the total length of the nucleotide sequence analyzed.

- 24 -

As used herein, “n” is the length of a repeat within said nucleotide sequence encoding the polypeptide consisting of proline, alanine and, optionally, serine. The length n is also the number of nucleotides of said repeat. Hence, n is the length of a repeat within the nucleotide sequence analyzed. By definition, the longest possible repeat may have a length that is by one nucleotide shorter than the total length of the nucleotide sequence analyzed (N_{tot}), i.e. $n = N_{\text{tot}} - 1$. On the other hand, the shortest length of a repeat considered for the NRS analysis is $n = 4$, which corresponds to the shortest stretch of nucleotides that is longer than a single triplet codon for an amino acid. As such codons occur multiple times just as a consequence of the encoded amino acid sequence they should not be regarded as nucleotide sequence repeats with regard to the technical problem to be solved.

The term “repeat” as used herein means that the nucleotide sequence comprises an identical contiguous sequence of nucleotides of a length n (i.e., the repeat) more than once. In other words, the nucleotide sequence comprises a contiguous part/stretch/sequence of a certain length of nucleotides in at least two or multiple copies. In other words, the term repeat refers to sequences of nucleotides of length n that are present in the nucleotide sequence more than once. It is contemplated herein that there may be just one type of repeat of a length n or there may be more than one different repeats of the same length n within the total length of the nucleotide sequence analyzed. Thus, a nucleotide sequence can, for example, have a repeat of a length n, whose occurrences all have the same sequence; it is understood that said repeat occurs at least twice but may also occur multiple times within the total length of the nucleotide sequence analyzed. Alternatively, there are repeats of the same length n that have different sequences, i.e., the repeats share the same length (n) but are not identical on the sequence level. In this instance, each different repeat sequence constitutes another type of repeat.

As used herein, “ $f_i(n)$ ” is the frequency of a repeat of length n. In other words, $f_i(n)$ is the number of occurrences of the repeat of length n. If there is only one type of repeat of length n, $k(n)$ is 1. Alternatively, if there are more than one different repeats of length n, $k(n)$ is the number of different sequences of said repeats of length n. In other words, if there is more than one repeat of length n with different sequences, $k(n)$ is the number of said different sequences of said repeats of length n. Thus, if there are two or more repeats of length n of different type, $k(n)$ is the number of said different sequences of said repeats of length n. Otherwise, if the repeats of a length n all have the same (i.e., identical) sequence, $k(n)$ is 1 for said repeat of length n.

For example, the frequency of a repeat of a length of 14 nucleotides ($n = 14$) may be five within a nucleotide sequence (of length N_{tot}). This can mean that the five repeats with a length of 14 nucleotides all have an identical sequence (belonging to the same type), i.e., the sequence of the

- 25 -

repeat occurs five times within the nucleotide sequence. In this case, $k(n)$ is 1 and $f_1(n) = 5$. In another scenario, the five repeats of 14 nucleotides length may have different sequences. In this scenario, it is envisaged that two of the five repeats share an identical sequence (constituting one type) and three of the five repeats share another identical sequence (constituting a second type), whereby the sequence of the former two repeats sharing one sequence and the sequence of the latter three repeats sharing another sequence are different from each other. Therefore, in this scenario, the number of said different sequences of length n is 2, i.e., $k(n)$ is 2, and $f_1(n) = 2$ and $f_2(n) = 3$. The index "i" in the term $f_i(n)$ can be understood to represent the type of repeat within a set of different repeats having the name length n .

The formula to determine the NRS comprises the sum of the squared length of repeats (n^2), wherein n is the index of summation, 4 is the lower bound of summation and $N_{\text{tot}} - 1$ is the upper bound of summation. Hence, the length of the repeat that is considered minimally is 4. A repeat of a length of 4 nucleotides includes all sequences longer than one amino acid codon triplet. The index, n , is incremented by 1 for each successive term, stopping when $n = N_{\text{tot}} - 1$.

Furthermore, the formula to determine the NRS comprises the square root of the sum of the frequencies of said repeats of length n ($f_i(n)$), wherein i is the index of the summation, $f_i(n)$ is an indexed variable representing each successive term in the series, 1 is the lower bound of summation and $k(n)$ is the upper bound of summation. The index, i , is incremented by 1 for each successive term, stopping when $i = k(n)$. Thus, if there is only one type of repeat of a length n , i.e. all repeats of length n have identical sequence, $k(n)$ is 1 for said repeat of length n and, instead of a sum, just the frequency $f_1(n)$ of this repeat of length n within the total length of the nucleotide sequence (N_{tot}) is analyzed.

For example, the following NRS-Calculator script may be used in accordance with the invention to determine an NRS:

```
import math
import sys
class NRSCalculator:
    def __init__(self):
        self.repeats = dict()
        self.sums = dict()
        self.seq = None
        self.range_min = None
        self.range_max = None
    def _match_at(self, row, column):
        return self.seq[row] == self.seq[column]
    def _get_repeats_at(self, row, column):
        length = 1
        search_row = row
```

- 26 -

```

search_column = column
while True:
    if not 0 <= search_row < len(self.seq):
        break
    if not 0 <= search_column < search_row:
        break
    if length > self.range_max:
        break
    if not self._match_at(search_row, search_column):
        break
    if length >= self.range_min:
        repeats =
self.repeats.setdefault(self.seq[row:row + length], set())
        repeats.add(row)
        repeats.add(column)
        search_row += 1
        search_column += 1
        length += 1
def _get_repeats(self):
    self.repeats = dict()
    for row in xrange(len(self.seq)):
        for column in xrange(row):
            self._get_repeats_at(row, column)
def _get_sums(self):
    self.sums = dict()
    for (seq, repeats) in self.repeats.iteritems():
        length = len(seq)
        self.sums[length] = self.sums.get(length, 0) +
len(repeats)
    def set_range(self, range_min, range_max):
        self.range_min = range_min
        self.range_max = range_max
    def set_sequence(self, seq):
        self.seq = seq
    def work(self):
        if not self.seq and not self.range_min and not
self.range_max:
            raise RuntimeError('Can not work without
initialization')
        self._get_repeats()
        self._get_sums()
    def print_repeats(self):
        print('Sequence (Length bp) : NumRepeats
(Positions)')
        for seq, repeats in sorted(self.repeats.iteritems(),
key=lambda t: len(t[0])):
            list = [seq, len(seq), len(repeats)]
            list.extend(map(lambda value: value + 1,
sorted(repeats)))
            print('%s Ntot = %u : %u (%s)' % (seq, len(seq),
len(repeats), ','.join(map(lambda value: str(value + 1),
sorted(repeats)))))
    def print_sums(self):

```

- 27 -

```

        print('Length\tFrequency')
        for item in self.sums.iteritems():
            print('%u\t%u' % item)
    def print_score(self):
        sum = 0
        for length, count in self.sums.iteritems():
            sum += (length ** 2) * math.sqrt(count)
        print('NRS = %.0f' % (sum / len(self.seq)))
def handle_sequence(finder, name, sequence):
    finder.set_range(4, len(sequence))
    finder.set_sequence(sequence)
    finder.work()
    print('%s: Ntot = %u' % (name, len(sequence)))
    #finder.print_repeats()
    finder.print_sums()
    finder.print_score()
if len(sys.argv) != 2:
    print('Usage: %s FILENAME' % sys.argv[0])
    sys.exit(1)
finder = NRSCalculator()
with open(sys.argv[1], 'r') as infile:
    name = 'Unnamed'
    seq = ''
    for line in infile:
        line = line.strip()
        if line.startswith('>'):
            if len(seq) > 0:
                handle_sequence(finder, name, seq)
            name = line
            seq = ''
            continue
        seq += line.upper()
    handle_sequence(finder, name, seq)

```

Furthermore, the invention also relates to a nucleotide sequence that comprises nucleotide repeats, wherein said repeats have a maximum length $n_{\max}$, wherein $n_{\max}$ is determined according to the formula

$$n_{\max} \leq 17 + \frac{N_{\text{tot}}}{600}$$

and wherein N_{tot} is the total length of said nucleotide sequence. The term “maximum length” or “maximal length” or “ $n_{\max}$ ” as used herein defines the number of nucleotides of the longest contiguous part/stretch/sequence of nucleotides that is present in at least two copies within said nucleotide sequence or nucleic acid molecule. In other words, the term “maximum length” or “maximal length” or “ $n_{\max}$ ” as used herein means that the nucleotide sequence of the nucleic acid molecule according to this invention has no repeats which are longer than this length.

- 28 -

It is demonstrated in the appended examples that exemplary nucleic acid molecules of the invention comprise only few repeats, e.g. of a length of 14 nucleotides; see e.g. appended Example 2. As explained above, the repeat analysis can be performed with any suitable tool such as the NRS analysis provided herein, manually or with the aid of generic software programs such as the dot plot analysis, for example using Visual Gene Developer (Jung (2011) loc. cit) or the Repfind tool (Betley (2002) loc. cit). A dot plot is a visual representation of the similarities between two sequences.

In the appended examples the nucleotide sequences provided herein were aligned to themselves. Each axis of a rectangular array represents one of the two (overall identical) nucleotide sequences to be compared. All positions from the first input sequence are compared with all positions from the second input sequence and scored, using a specified substitution matrix. This produces a matrix of scores from which local regions of similarity/identity (corresponding to diagonals in the dot plot) are identified. A repeat window/threshold of user-specified length is moved along all possible diagonals. Each position in the repeat window/threshold corresponds to a pair-wise score from the scoring matrix. The score for the entire window is the sum of the scores for individual positions within it. If the window score is above the user-defined threshold, then a line is plotted in the dot plot corresponding to the window (see e.g., <http://emboss.sourceforge.net/apps/cvs/emboss/apps/dotmatcher.html>).

The dot plot analysis is employed in the appended examples, e.g. Example 3, in order to analyze the repetitiveness of the nucleotide sequences of the nucleic acid molecules. It is proven therein that the inventive nucleic acid molecules, e.g., SEQ ID NO: 36 (termed PA#3b(200) herein) or SEQ ID NO: 28 (termed PA#1b herein), have only a few scattered 14 nucleotide repeats within the entire length of 600 nucleotides in case a repeat window/threshold of 14 is applied. An increase of the repeat window/threshold of 14 by one nucleotide, i.e., a repeat window/threshold of 15 nucleotides, reveals no further repeats within the whole nucleotide sequence analyzed (see Figure 2 and Example 3). This means that the maximal length of the repeat within the entire nucleic acid sequence investigated has a length of 14 nucleotides (also including shorter repeats). Generally, one can assume that the shorter such nucleotide repeats are, the less they play a detrimental role for genetic stability.

Yet, nucleic acid molecules encoding proline/alanine-rich sequences disclosed in the prior art show longer and/or more repeat sequences if a repeat window/threshold of, for example, 14 or 15 nucleotides is applied as used in the appended Examples (see Example 3). For example, the proline/alanine-rich sequence disclosed in WO 2011/144756 (SEQ ID NO: 15; termed PA#3a(200) herein) possesses multiple nucleotide repeats of length 60, which are consecutive and even overlapping and thus give rise to the large number of long parallel lines. Further

- 29 -

nucleic acid molecules disclosed in the prior art such as PAS#1a(600) as shown in SEQ ID NO: 12 and as disclosed in WO 2008/155134, an $[(AP)_5]_n$ multimer as shown in SEQ ID NO: 16 and as disclosed in US2006/0252120, or the large tegument protein of *Macacine herpesvirus 1* as shown in SEQ ID NO: 18 and as disclosed in GenBank entry AAP41454.1 also document the highly repetitive nature of these nucleotide sequences.

This is in stark contrast to the nucleic acid molecules of the present invention, which demonstrate low repetitiveness on the nucleotide sequence level. When, for example, the number of repeats of the prior art sequence PA#3a(200) (SEQ ID NO: 15) is compared, using a repeat window/threshold of 14 nucleotides, to the number of repeats of a low repetitive nucleotide sequence of the invention, PA#3b(200) (SEQ ID NO: 36), it can be observed that the sequence of the invention has 29 repeats (per 600 nucleotide residues) compared to many more than 100 repeats (per 600 nucleotide residues) of the prior art sequence. If a window of 15 nucleotides is applied, the exemplary analyzed nucleic acid molecule or the nucleotide sequence of the invention does not comprise any repeats. In contrast, the nucleic acid molecules of the prior art reveal numerous repeats with a length of 15 nucleotides and more. As proven in the appended examples, a repeat window/threshold of 14 to 20, e.g. 14, 15, 16, 17, 18, 19 or 20 nucleotides, in particular, is suitable to confirm that a nucleic acid molecule of the invention has a low repetitive nucleotide sequence. It is appreciated that the length of the repeat window/threshold inversely correlates with the number of repeats that will be identified in a specific nucleotide sequence. For example, if the repeat window/threshold is "1" the number of repeats may equal the number of all nucleotide residues (A, T, G and/or C) in a nucleotide sequence (provided that each type of nucleotide occurring in the analyzed sequence is present in at least two copies). If the length of the repeat window/threshold increases, the number of repeats in the analyzed nucleotide sequence will decrease in a sequence-specific manner. Consequently, also the "low repetitive" nucleotide sequences of the invention can contain repeats; yet, compared to prior art sequences these are shorter and fewer if the same parameters are applied for analysis.

Therefore, the nucleic acid molecule(s) provided herein comprise(s) repeats of a maximum length of 14, 15, 16, 17, about 18, about 19, about 20, about 21, about 25, about 30, about 35, about 40, about 45, about 50 or about 55 nucleotides. Furthermore, the nucleic acid molecule provided herein comprises repeats of a maximum length of about 17 nucleotides to a maximum length of about 55 nucleotides. In the context of the maximum length of the repeat, the term "about" means that the maximal repeat length is +/- 4 nucleotides of the indicated repeat length. In other words, in this context, "about" refers to a range, wherein the length of the repeat can be 4 nucleotides longer or shorter than the indicated repeat length. For example, a maximum repeat length of about 55 nucleotides refers to a nucleotide sequence comprising

- 30 -

repeats of a maximum length of 51 to 59 nucleotides. Furthermore, a maximum repeat length of about 17 nucleotides refers to a nucleotide sequence comprising repeats of a maximum length of 13 to 21 nucleotides.

Furthermore, the invention relates to a nucleic acid molecule comprising repeats of a maximum length of 59 nucleotides, preferably 54 nucleotides, more preferably 50 nucleotides, more preferably 48 nucleotides, more preferably 40 nucleotides, more preferably 36 nucleotides, more preferably 35 nucleotides, more preferably 30 nucleotides, more preferably 25 nucleotides, more preferably 24 nucleotides, more preferably 21 nucleotides, more preferably 20 nucleotides, more preferably 19 nucleotides, more preferably 18 nucleotides, more preferably 16 nucleotides, more preferably 15 nucleotides, more preferably 14 nucleotides, or most preferably 17 nucleotides.

As used herein, a nucleic acid molecule comprising repeats of a maximum length of, for example, 17 nucleotides relates to a nucleic acid molecule comprising repeat sequences with lengths of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 or 17 nucleotides. Likewise, the nucleic acid molecule comprising repeats of a maximum length of 14 nucleotides relates to a nucleic acid molecule comprising repeats up to 14 nucleotides, i.e., " ≤ 14 nucleotides" or "1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 or 14 nucleotides". In other words, the nucleic acid molecule comprising repeats of a maximum length of 14 nucleotides does not comprise repeats that are longer than 14 nucleotides, i.e., " >14 nucleotides".

The invention furthermore relates to a low repetitive nucleic acid molecule, wherein low repetitive means that a nucleotide sequence stretch which repeatedly occurs within a longer nucleotide sequence encoding a repetitive amino acid sequence corresponds to maximally 0.05%, preferably 0.1%, more preferably 0.5%, more preferably 1%, more preferably 2%, more preferably 3%, more preferably 4%, more preferably 5%, more preferably 6%, more preferably 7%, more preferably 8%, more preferably 9%, more preferably 10%, more preferably 15%, more preferably 20%, more preferably 25%, more preferably 30%, more preferably 40%, or most preferably 50 % of the length of the nucleotide sequence that encodes the repeated amino acid sequence stretch. In other words, the nucleic acid of the present invention comprises repeats, wherein said repeats have a maximum length corresponding to 0.05%, preferably 0.1%, more preferably 0.5%, more preferably 1%, more preferably 2%, more preferably 3%, more preferably 4%, more preferably 5%, more preferably 6%, more preferably 7%, more preferably 8%, more preferably 9%, more preferably 10%, more preferably 15%, more preferably 20%, more preferably 25%, more preferably 30%, more preferably 40%, or most preferably 50 % of the length of said nucleotide sequence encoding the amino acid sequence repeat in the polypeptide consisting of proline, alanine and, optionally, serine.

It is evident that the present invention does not only provide novel and inventive nucleotide sequences encoding PA-rich sequences and having a length of at least 300 nucleotides (corresponding to 100 amino acid residues) but the present invention also provides for selection means and methods for recombinant and/or synthetic nucleic acid molecules leading to genetically stable sequences and/or allowing convenient cloning, sequencing and/or amplification. This selection method is based on the NRS provided herein and/or the maximal repeat length described herein. Accordingly, the present invention relates to a method for selecting a genetically stable nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein said nucleotide sequence has a length of at least 300 nucleotides, the method comprising a step of selecting a nucleic acid molecule comprising a nucleotide sequence having a Nucleotide Repeat Score (NRS) lower than 50,000, wherein said Nucleotide Repeat Score (NRS) is determined according to the formula provided herein above. Furthermore, the present invention relates to a method for selecting a genetically stable nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, the method comprising a step of selecting said nucleotide sequence comprising repeats having a maximum length of $n_{\max}$. Furthermore, the present invention relates to a method for selecting a genetically stable nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine, the method comprising a step of selecting said nucleotide sequence comprising repeats of a maximum length of about 17 nucleotides to a maximum length of about 55 nucleotides. The embodiments and the parameters of the NRS formula and the low repetitive nucleotide sequences provided herein above in the context of the inventive nucleic acids encoding PA-rich sequences apply, *mutatis mutandis*, for the herein provided and illustrated selection methods for genetically stable nucleic acid molecules encoding PA-rich sequences as well as for the further embodiments.

Herein, it was surprisingly found and demonstrated that the inventive nucleic acid molecules have an enhanced *in vivo* stability. As shown in the appended illustrative examples, e.g., Example 6, the nucleic acid molecules provided herein are more stable than nucleic acid molecules comprising more repetitive nucleotide sequences. Example 6 compares nucleic acid molecules encoding polypeptides consisting of proline, alanine and serine that differ in the repetitiveness of the nucleotide sequences. The inventive nucleic acid molecules comprising low repetitive nucleotide sequences provided herein (see also Fig. 2B), e.g., as given in SEQ ID NO: 50, have a higher genetic stability compared to highly repetitive nucleotide sequences such as given in the prior art, e.g., SEQ ID NO: 51. Thus, the term “*in vivo* stability” as used herein refers particularly to “genetic stability”. The term “genetic stability” as used herein means that

the nucleic acid is stably maintained in the host cell and that the sequence is not genetically modified, e.g., by mutation, insertion or deletion. As used herein, mutations refer to changes in the nucleotide sequence, e.g., substitutions, deletions, insertions or extensions. It is shown in the appended examples that nucleic acid molecules comprising highly repetitive nucleotide sequences are prone to deletions or shortening; see, e.g., Example 6. The nucleic acid molecules comprising highly repetitive nucleotide sequences/repetitive gene cassettes have thus an inferior genetic stability, which may arise from recombination events during cell division, for example. The parameters NRS or $n_{\max}$ defined herein provide objective means to distinguish highly repetitive nucleotide sequences such as those from the prior art, having high numbers for NRS and/or $n_{\max}$, from low repetitive nucleotide sequences of the invention, which have low numbers for NRS and/or $n_{\max}$.

The nucleic acid molecules provided herein encode polypeptides comprising repetitive amino acid sequences. Notably, the encoded repetitive amino acid sequence may also be part/fragment of a fusion protein. For example, the nucleic acid molecule disclosed herein can encode a polypeptide consisting of proline/alanine-rich repeats, e.g. repeats consisting of proline, alanine and, optionally, serine. The proline/alanine-rich polypeptide can form a random coil. In certain aspects, the nucleic acid molecule disclosed herein encodes a polypeptide consisting of proline, alanine and serine, wherein said polypeptide forms a random coil. In certain aspects, the nucleic acid molecule disclosed herein encodes a polypeptide consisting of proline and alanine, wherein said polypeptide forms a random coil.

In particular, the nucleotide sequence encodes a polypeptide consisting of proline, alanine and, optionally, serine. This encoded polypeptide forms a random coil. The nucleotide sequence is comprised in the nucleic acid molecule provided herein. Thus, in certain aspects, the nucleic acid molecule disclosed herein encodes a polypeptide comprising a repetitive amino acid sequence and consisting of proline, alanine and, optionally, serine, wherein said polypeptide forms a random coil. In certain aspects, the nucleic acid molecule disclosed herein encodes a polypeptide comprising a repetitive amino acid sequence and consisting of proline, alanine and serine, wherein said polypeptide forms a random coil. In certain aspects, the nucleic acid molecule disclosed herein encodes a polypeptide comprising a repetitive amino acid sequence and consisting of proline and alanine, wherein said polypeptide forms a random coil.

As used herein, the term "random coil" relates to any conformation of a polymeric molecule, including amino acid polymers, in particular polypeptides made of L-amino acids, in which the individual monomeric elements that form said polymeric structure are essentially randomly oriented towards the adjacent monomeric element or elements while still being chemically linked. In particular, the encoded polypeptide or amino acid polymer adopting/having/forming

"random coil conformation" substantially lacks a defined secondary and tertiary structure. The nature of the encoded polypeptide random coils and their methods of experimental identification are known to the person skilled in the art and have been described in the scientific literature (Cantor (1980) Biophysical Chemistry, 2nd ed., W. H. Freeman and Company, New York; Creighton (1993) Proteins – Structures and Molecular Properties, 2nd ed., W. H. Freeman and Company, New York; Smith (1996) Fold. Des. 1:R95-R106) and patent literature, e.g., WO2011/144756 and WO2008/155134.

The encoded random coil polypeptides of the present invention adopt/form a random coil conformation, for example, in aqueous solution and/or at physiological conditions. The term "physiological conditions" is known in the art and relates to those conditions in which proteins usually adopt their native, folded conformation. More specifically, the term "physiological conditions" relates to the environmental biophysical parameters as they are typically valid for higher forms of life and, particularly, for mammals, most preferably human beings. The term "physiological conditions" may relate to the biochemical and biophysical parameters as they are normally found in the body, in particular in body fluids, of mammals and in particular in humans. Said "physiological conditions" may relate to the corresponding parameters found in the healthy body as well as the parameters found under disease conditions or in human patients. For example, a sick mammal or human patient may have a higher, yet "physiological" body temperature (i.e., temperature condition) when said mammal or said human suffers from fever. With respect to "physiological conditions" at which proteins adopt their native conformation/state, the most important parameters are temperature (37 °C for the healthy human body), pH (7.35–7.45 for human blood), osmolarity (280–300 mmol/kg H₂O), and, if necessary, general protein content (66–85 g/l serum).

Yet, the person skilled in the art is aware that at physiological conditions these parameters may vary, e.g. the temperature, pH, osmolarity, and protein content may be different in given body or tissue fluids such as blood, liquor cerebrospinalis, peritoneal fluid and lymph (Klinke (2005) Physiologie, 4th edition., Georg Thieme Verlag, Stuttgart). For example, in the liquor cerebrospinalis the osmolarity may be around 290 mmol/kg H₂O and the protein concentration may be between 0.15 g/l and 0.45 g/l while in the lymph the pH may be around 7.4 and the protein content may be between 3 g/l and 5 g/l. When determining whether an encoded polypeptide/amino acid sequence forms/adopts random coil conformation under experimental conditions, the biophysical parameters such as temperature, pH, osmolarity and protein content may be different from the physiological conditions normally found *in vivo*. Temperatures between 1 °C and 42 °C or preferably 4 °C to 25 °C may be considered useful to test and/or verify the biophysical properties and biological activity of a protein under physiological conditions *in vitro*.

Several buffers, which may include solvents and/or excipients for pharmaceutical compositions, are considered to represent “physiological solutions” / “physiological conditions” *in vitro*, in particular, in experimental settings, for example in the context of CD measurements or other methods that allow the person skilled in the art to determine the structural properties of a protein/amino acid sequence. Examples of such buffers are, e.g., phosphate-buffered saline (PBS, e.g.: 115 mM NaCl, 4 mM KH_2PO_4 , 16 mM Na_2HPO_4 pH 7.4), Tris buffers, acetate buffers, citrate buffers or similar buffers. Generally, the pH of a buffer representing “physiological solution conditions” should lie in a range from 6.5 to 8.5, preferably in a range from 7.0 to 8.0, most preferably in a range from 7.2 to 7.7, and the osmolarity should lie in a range from 10 to 1000 mmol/kg H_2O , more preferably in a range from 50 to 500 mmol/kg H_2O and most preferably in a range from 200 to 350 mmol/kg H_2O . Optionally, the protein content of a physiological solution may lie in a range from 0 to 100 g/l, neglecting the investigated protein adopting random coil conformation itself; furthermore, typical stabilizing proteins may be present, for example human or bovine serum albumin.

The polypeptides encoded by the nucleic acid molecules of the invention not only form random coil conformation under physiological conditions but, more generally, in aqueous solution; e.g., c.f. WO2011/144756. The term “aqueous solution” is well known in the art. An “aqueous solution” may be a solution with a water (H_2O) content of at least about 20 %, of at least about 30 %, of at least about 40 %, of at least about 50 %, of at least about 60%, of at least about 70%, of at least about 80% or of at least about 90% H_2O (weight/weight). Accordingly, the encoded polypeptides provided in the present invention may form random coil conformation in aqueous solution, possibly containing other miscible solvents, or in aqueous dispersions with a wider range of temperatures, pH values, osmolarities or protein content. This is particularly relevant for applications of the random coil polypeptide outside medical therapy or *in vivo* diagnostics, for example in cosmetics, nutrition or food technology.

It is also envisaged in the context of this invention that the random coil conformation of the encoded polypeptide is maintained in pharmaceutical compositions like liquid pharmaceuticals/biologicals or lyophilized pharmaceutical compositions. This is particularly important in the context of the encoded biologically active proteins or the drug conjugates provided herein comprising, *inter alia*, the random coil polypeptide. Preferably, “physiological conditions” are to be used in corresponding buffer systems, solvents and/or excipients. Yet, for example, in lyophilized or dried compositions (like, e.g., pharmaceutical compositions/biologicals), it is envisaged that the random coil conformation of the herein provided encoded random coil polypeptide may transiently not be present and/or cannot be detected. However, said encoded random coil polypeptide will adopt/form its random coil

again after reconstitution in corresponding buffers/solutions/excipients/solvents or after administration to the body of a patient or of an animal.

In certain aspects of the present invention, the nucleic acid molecules disclosed herein encode polypeptides that (mainly or exclusively) consist of proline, alanine and, optionally, serine, wherein no more than 9 consecutive amino acid residues are identical. Such encoded polypeptides form a random coil. In a certain aspect, the encoded amino acid sequences/polypeptides adopting random coil conformation may comprise a plurality of amino acid repeats, wherein said "amino acid repeats" mainly or exclusively consist of proline, alanine and, optionally, serine amino acid residues, wherein no more than 9 consecutive amino acid residues are identical. In a certain aspect, the encoded amino acid sequences/polypeptides adopting random coil conformation (the random coil polypeptide as defined herein) may comprise a plurality of amino acid repeats, wherein said "amino acid repeats" mainly or exclusively consist of proline, alanine and serine amino acid residues, wherein no more than 9 consecutive amino acid residues are identical. In a certain aspect, the encoded amino acid sequences/polypeptides adopting random coil conformation may comprise a plurality of amino acid repeats, wherein said "amino acid repeats" mainly or exclusively consist of proline and alanine amino acid residues, wherein no more than 9 consecutive amino acid residues are identical.

In preferred aspects, the nucleic acid molecule disclosed herein encodes a polypeptide comprising a repetitive amino acid sequence with a plurality of amino acid repeats, wherein no more than 8 consecutive amino acid residues are identical and wherein said polypeptide forms a random coil. Preferably, the nucleic acid molecule disclosed herein encodes a polypeptide comprising a repetitive amino acid sequence with a plurality of amino acid repeats, wherein no more than 7 consecutive amino acid residues are identical and wherein said polypeptide forms a random coil. More preferably, the nucleic acid molecule disclosed herein encodes a polypeptide comprising a repetitive amino acid sequence with a plurality of amino acid repeats, wherein no more than 6 consecutive amino acid residues are identical and wherein said polypeptide forms a random coil. Particularly preferably, the nucleic acid molecule disclosed herein encodes a polypeptide comprising a repetitive amino acid sequence with a plurality of amino acid repeats, wherein no more than 5 consecutive amino acid residues are identical and wherein said polypeptide forms a random coil. More particularly preferably, the nucleic acid molecule disclosed herein encodes a polypeptide comprising a repetitive amino acid sequence with a plurality of amino acid repeats, wherein no more than 4 consecutive amino acid residues are identical and wherein said polypeptide forms a random coil. Most preferably, the nucleic acid molecule disclosed herein encodes a polypeptide comprising a repetitive amino acid

sequence with a plurality of amino acid repeats, wherein no more than 3 consecutive amino acid residues are identical and wherein said polypeptide forms a random coil.

A non-limiting example of an amino acid repeat consisting exclusively of proline, alanine and serine residues is provided herein below; see, e.g. SEQ ID NO: 6. A non-limiting example of an encoded amino acid repeat consisting exclusively of proline and alanine residues is provided herein below; see, e.g. SEQ ID NO: 8. The encoded polypeptide may comprise multiple copies of the same sequence or different sequences.

The nucleic acid molecule disclosed herein encodes a polypeptide consisting mainly or exclusively of the three amino acid residues proline (Pro, P), alanine (Ala, A) and, optionally, serine (Ser, S). The term “optionally” as used herein means that the encoded polypeptide of the present invention either consists mainly or exclusively of proline, alanine and serine or consists mainly or exclusively of proline and alanine. The encoded polypeptide consisting mainly or exclusively of the three amino acid residues proline, alanine and serine is referred to herein as “PAS” polypeptide. The encoded polypeptide consisting mainly or exclusively of the two amino acid residues proline and alanine is referred to herein as “PA” polypeptide. A non-limiting example of an encoded polypeptide consisting of proline and alanine is given in SEQ ID NO: 8. A non-limiting example of an encoded polypeptide consisting of proline, alanine and serine is given in SEQ ID NO: 6. The term “mainly” as used herein means that preferably at least about 90% or at least about 95% of the encoded amino acids are proline, alanine and, optionally, serine, whereby proline, alanine and serine in sum constitute the majority but may not be the only amino acid residues; therefore, the encoded amino acid sequences are not necessarily 100 % proline, alanine and, optionally, serine. Hence, the encoded polypeptides/amino acid sequences may also comprise other amino acids than proline, alanine and, optionally, serine as minor constituents as long as the amino acid sequence forms/adopts/has the random coil conformation. Such a random coil conformation can be easily determined by means and methods described herein. Accordingly, the present invention relates in one embodiment to a nucleic acid molecule that encodes a random coil polypeptide whereby the amino acid sequence consists mainly of proline, alanine and, optionally, serine.

In case the encoded polypeptide consists of proline and alanine, said proline residues constitute more than about 10% and less than about 75 % of said encoded amino acid sequence. Accordingly, the encoded random coil polypeptide consists mainly of proline and alanine, wherein the proline residues constitute more than about 10 % and less than 75 % of the amino acid sequence. The alanine residues comprise the remaining at least 25 % to 90 % of said amino acid sequence.

- 37 -

Preferably, the encoded amino acid sequence comprises more than about 10%, preferably more than about 12%, more preferably more than about 14%, 18%, 20%, more preferably more than about 22%, 23%, 24%, or 25%, more preferably more than about 27%, 29%, or 30%, more preferably more than about 32%, 33%, or 34% and most preferably more than about 35% proline residues. The amino acid sequence preferably comprises less than about 75%, more preferably less than 70%, more preferably less than 65%, more preferably less than 60%, more preferably less than 55%, more preferably less than 50% proline residues, wherein the lower values are preferred. Even more preferably, the amino acid sequence comprises less than about 48%, 46%, 44%, 42% proline residues. More preferred are amino acid sequences comprising less than about 41%, 40%, 39%, 38%, 37% or 36% proline residues, whereby lower values are preferred. More preferred are amino acid sequences comprising less than about 34%, 32%, or 30%. More preferred are amino acid sequences comprising less than about 28%, 26% or 25%. Most preferably, the amino acid sequences comprise less than about 35% proline residues.

Vice versa, the amino acid sequence preferably comprises less than about 90 %, more preferably less than 88 %, 86 %, 84 %, 82 % or 80 % alanine residues, wherein the lower values are preferred. More preferably, the amino acid sequence comprises less than about 79 %, 78 %, 77 %, 76 % alanine residues, whereby lower values are preferred. More preferably, the amino acid sequence comprises less than about 74 %, 72 %, or 70 % alanine residues, whereby lower values are preferred. More preferably, the amino acid sequence comprises less than about 69 %, 67 %, or 65 % alanine residues, whereby lower values are preferred. Most preferably, the amino acid sequence comprises less than about 75% alanine residues. Also preferred herein is an amino acid sequence comprising more than about 25%, preferably more than about 30%, more preferably more than about 35%, more preferably more than about 40%, more preferably more than about 45%, more preferably more than about 50%, more preferably more than about 52%, 54%, 56%, 58% or 59% alanine residues, wherein the higher values are preferred. Even more preferably, the amino acid sequence comprises more than about 60 %, 61 %, 62 %, 63 % or 64 % alanine residues. More preferably, the amino acid sequence comprises more than about 66%, 67%, 69%, or 70% alanine residues. More preferably, the amino acid sequence comprises more than about 72%, 74%, or 75%, alanine residues. Most preferably the amino acid sequence comprises more than about 65 % alanine residues.

Accordingly, the random coil polypeptide may comprise an amino acid sequence consisting of about 25 % or 30% proline residues and about 75 % or 70 %, respectively, alanine residues. Alternatively, the random coil polypeptide may comprise an amino acid sequence consisting of about 35 % proline residues and about 65 % alanine residues. The term "about X %" as used herein above is not limited to the concise number of the percentage, but also comprises values of 10 % to 20 % additional or 10 % to 20 % less residues. For example, the term 10 % may also

- 38 -

relate to 11 % or 12 % and to 9 % or 8 %, respectively.

In case the encoded polypeptide consists of proline, alanine and serine, said proline residues constitute more than about 4% and less than about 40% of said encoded amino acid sequence. The alanine and the serine residues constitute the remaining amount of said amino acid sequence.

Preferably, the encoded amino acid sequence comprises more than about 4 %, preferably more than about 6 %, more preferably more than about 10 %, more preferably more than about 15 %, more preferably more than about 20 %, more preferably more than about 22%, 23% or 24%, more preferably more than about 26%, 29%, or 30%, more preferably more than about 31%, 32%, 33%, 34% or 35% and most preferably more than about 25 % proline residues. The encoded amino acid sequence preferably comprises less than about 40 %, more preferably less than 38%, 35%, 30%, 26% proline residues, wherein the lower values are preferred.

The encoded amino acid sequence preferably comprises less than about 95 %, more preferably less than 90%, 86%, 84%, 82% or 80% alanine residues, wherein the lower values are preferred. More preferably, the encoded amino acid sequence comprises less than about 79%, 78%, 77%, 76% alanine residues, whereby lower values are preferred. More preferably, the encoded amino acid sequence comprises less than about 75%, 73%, 71%, or 70% alanine residues, whereby lower values are preferred. More preferably, the encoded amino acid sequence comprises less than about 69%, 67%, 66%, or 65% alanine residues, whereby lower values are preferred. More preferably, the encoded amino acid sequence comprises less than about 64%, 63%, 62%, or 60% alanine residues, whereby lower values are preferred. More preferably, the encoded amino acid sequence comprises less than about 59%, 57%, 56%, or 55% alanine residues, whereby lower values are preferred. More preferably, the encoded amino acid sequence comprises less than about 54%, 53%, or 51%, alanine residues, whereby lower values are preferred. Most preferably, the encoded amino acid sequence comprises less than about 50 % alanine residues.

Also preferred herein is an encoded amino acid sequence comprising more than about 10 %, preferably more than about 15%, 17%, 19%, or 20%, more preferably more than about 22%, 24%, or 25%, , more preferably more than about 27%, 29%, or 30%, more preferably more than about 32%, 34% or 35%, more preferably more than about 37%, 39%, or 40%, more preferably more than about 42%, 44% or 45%, more preferably more than about 46%, 47% or 49% alanine residues, wherein the higher values are preferred. Most preferably, the encoded amino acid sequence comprises more than about 50 alanine residues. As mentioned above, the serine residues comprise the remaining amount of said amino acid sequence.

Accordingly, the encoded random coil polypeptide may comprise an amino acid sequence consisting of about 35 % proline residues, about 50 % alanine and 15 % serine residues. Exemplary nucleotide sequences and the encoded polypeptides thereof can be found in Table 1. The term "about X %" as used herein above is not limited to the concise number of the percentage, but also comprises values of 10 % to 20 % additional or 10 % to 20 % less residues. For example, the term 10 % may also relate to 11 % or 12 % or to 9 % and 8 %, respectively.

However, as mentioned above and further detailed herein below said encoded random coil polypeptide, and, in particular, the amino acid sequence may also comprise additional amino acids differing from proline, alanine and, optionally, serine as minor constituents. As already discussed herein above, said minor constituent(s), i.e. amino acid(s) different from proline, alanine or, optionally, serine, may comprise less than about 10 % or less than about 5 % of the encoded random coil polypeptide of this invention.

The skilled person is aware that the encoded amino acid sequence/polypeptide may also form random coil conformation when other residues than proline, alanine and, optionally, serine are comprised as a minor constituent in said amino acid sequence/polypeptide (polypeptide fragment). The term "minor constituent" as used herein means that maximally 5 % or maximally 10 % amino acid residues are different from proline, alanine or serine in the encoded random coil polypeptides of this invention. This means that maximally 10 of 100 amino acids may be different from proline, alanine and, optionally, serine, preferably maximally 8 %, i.e. maximally 8 of 100 amino acids may be different from proline, alanine and, optionally, serine, more preferably maximally 6%, i.e. maximally 6 of 100 amino acids may be different from proline, alanine and, optionally, serine, even more preferably maximally 5%, i.e. maximally 5 of 100 amino acids may be different from proline, alanine and, optionally, serine, particularly preferably maximally 4%, i.e. maximally 4 of 100 amino acids may be different from proline, alanine and, optionally, serine, more particularly preferably maximally 3%, i.e. maximally 3 of 100 amino acids may be different from proline, alanine and, optionally, serine, even more particularly preferably maximally 2%, i.e. maximally 2 of 100 amino acids may be different from proline, alanine and, optionally, serine and most preferably maximally 1%, i.e. maximally 1 of 100 of the amino acids that are comprised in the random coil polypeptide may be different from proline, alanine and, optionally, serine. Said amino acids different from proline, alanine and, optionally, serine may be selected from the group consisting of Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Thr, Trp, Tyr, and Val, including posttranslationally modified amino acids or non-natural amino acids (see, e.g., Budisa (2004) *Angew Chem Int Ed Engl* 43:6426-6463; Young (2010) *J Biol Chem*

- 40 -

285:11039-11044; Liu (2010) *Annu Rev Biochem* 79:413-444; Wagner (1983) *AngewChem Int Ed Engl* 22:816-828; Walsh (2010) *Drug Discov Today* 15: 773-780. In certain cases PA-rich sequences can also comprise Ser as a minor constituent. For example, in case the encoded random coil polypeptide consists of proline and alanine, serine can also be considered as minor constituent.

Generally, it is preferred herein that these “minor” amino acids (other than proline, alanine and, optionally, serine) are not present in the encoded random coil polypeptide as described herein or the encoded random coil polypeptide as part/fragment of a fusion protein. In accordance with the invention, the encoded random coil polypeptide / amino acid sequence may, in particular, consist exclusively of proline, alanine and, optionally, serine residues (i.e. no other amino acid residues are present in the encoded random coil polypeptide or in the amino acid sequence).

In the context of the present invention, the nucleic acid molecule encoding the polypeptide consisting of proline, alanine and, optionally, serine or the nucleotide sequence encoding the polypeptide consisting of proline, alanine and, optionally, serine may consist of at least 300 nucleotides. However, it is evident for a person skilled in the art that the length of the nucleotide sequence provided herein is not limited as long as the encoded polypeptide forms a random coil. The appended examples document that the nucleic acid molecules provided herein comprising low repetitive nucleotide sequences can surprisingly be synthesized regardless of their lengths. It is proven herein that, for example, nucleotide sequences having a length of about 600 nucleotides can be synthesized. Such nucleotide sequences can be employed to assemble even longer nucleotide sequences. In other words, these nucleotide sequences can be employed as units/modules/building blocks to combine/assemble longer nucleotide sequences encompassed in the nucleic acid molecule of the invention. In one embodiment, a single nucleotide sequence unit/module/building block also corresponds to a nucleic acid molecule of the invention.

According to this invention, identical or non-identical nucleotide sequence units/modules/building blocks can be combined with each other as long as the assembled nucleotide sequence encodes a polypeptide consisting of proline, alanine and, optionally, serine. Furthermore, according to this invention, identical or non-identical nucleotide sequence units/modules/building blocks can be combined with each other as long as the assembled nucleotide sequence encoding a polypeptide with repetitive amino acid sequence forms a random coil. As mentioned above, it is particularly advantageous to assemble the nucleic acid molecules provided herein comprising low repetitive nucleotide sequences by employing these units/modules/building blocks. It is documented herein below that long nucleotide sequences with at least a length of 2400 nucleotides can be assembled. The inventive nucleotide

- 41 -

sequences can be combined with each other or with additional non-identical nucleotide sequences. Thus, the nucleic acid molecule provided herein may be assembled from identical or non-identical nucleotide sequences, wherein said nucleotide sequences are low repetitive nucleotide sequences. In certain aspects, the nucleic acid molecule provided herein comprises at least one nucleotide sequence selected from the group consisting of SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 26, SEQ ID NO: 27, SEQ ID NO: 28, SEQ ID NO: 29, SEQ ID NO: 30, SEQ ID NO: 31, SEQ ID NO: 32, SEQ ID NO: 33, SEQ ID NO: 34, SEQ ID NO: 35, SEQ ID NO: 36, SEQ ID NO: 37, SEQ ID NO: 87, SEQ ID NO: 88, SEQ ID NO: 89, SEQ ID NO: 90, SEQ ID NO: 91, SEQ ID NO: 92, SEQ ID NO: 93, SEQ ID NO: 94, SEQ ID NO: 95, SEQ ID NO: 96, SEQ ID NO: 97, SEQ ID NO: 98, SEQ ID NO: 99, SEQ ID NO: 100, SEQ ID NO: 101, SEQ ID NO: 102, SEQ ID NO: 103, SEQ ID NO: 104, SEQ ID NO: 105, SEQ ID NO: 106, SEQ ID NO: 107, SEQ ID NO: 108, SEQ ID NO: 109, SEQ ID NO: 110, SEQ ID NO: 111, SEQ ID NO: 112, SEQ ID NO: 113, SEQ ID NO: 114, SEQ ID NO: 115, SEQ ID NO: 116, SEQ ID NO: 117, SEQ ID NO: 118, SEQ ID NO: 119, SEQ ID NO: 120, SEQ ID NO: 121, SEQ ID NO: 122, SEQ ID NO: 123, SEQ ID NO: 124, SEQ ID NO: 125, SEQ ID NO: 126, SEQ ID NO: 127, SEQ ID NO: 128, SEQ ID NO: 129, SEQ ID NO: 130, SEQ ID NO: 131, SEQ ID NO: 132, SEQ ID NO: 133, SEQ ID NO: 134, SEQ ID NO: 135, SEQ ID NO: 136, SEQ ID NO: 137, SEQ ID NO: 138, SEQ ID NO: 139, SEQ ID NO: 140, SEQ ID NO: 141, SEQ ID NO: 142, SEQ ID NO: 143, SEQ ID NO: 144, SEQ ID NO: 145, SEQ ID NO: 146, SEQ ID NO: 147, SEQ ID NO: 148, SEQ ID NO: 149, SEQ ID NO: 150, SEQ ID NO: 151, SEQ ID NO: 152, SEQ ID NO: 192, SEQ ID NO: 193, SEQ ID NO: 194 and SEQ ID NO: 195. As these exemplary nucleotide sequences can be employed to assemble longer nucleotide sequences, these nucleotide sequences can be termed units or modules or building blocks. Thus, it is understood herein that the nucleic acid molecules of the invention may comprise a plurality of these nucleotide modules or nucleotide sequences that are assembled into a longer nucleotide sequence, wherein said longer nucleotide sequence itself is a low repetitive nucleotide sequence as described herein above. The person skilled in the art understands that the nucleic acid molecule of the invention may also comprise fragments of the given nucleotide sequence modules. In other words, the nucleic acid molecule provided herein comprises or is at least one nucleotide sequence, or fragment thereof, selected from the group consisting of SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 26, SEQ ID NO: 27, SEQ ID NO: 28, SEQ ID NO: 29, SEQ ID NO: 30, SEQ ID NO: 31, SEQ ID NO: 32, SEQ ID NO: 33, SEQ ID NO: 34, SEQ ID NO: 35, SEQ ID NO: 36, SEQ ID NO: 37, SEQ ID NO: 87, SEQ ID NO: 88, SEQ ID NO: 89, SEQ ID NO: 90, SEQ ID NO: 91, SEQ ID NO: 92, SEQ ID NO: 93, SEQ ID NO: 94, SEQ ID NO: 95, SEQ ID NO: 96, SEQ ID NO: 97, SEQ ID NO: 98, SEQ ID NO: 99, SEQ ID NO: 100, SEQ ID NO: 101, SEQ ID NO: 102, SEQ ID NO: 103, SEQ ID NO: 104, SEQ ID NO: 105,

- 42 -

SEQ ID NO: 106, SEQ ID NO: 107, SEQ ID NO: 108, SEQ ID NO: 109, SEQ ID NO: 110, SEQ ID NO: 111, SEQ ID NO: 112, SEQ ID NO: 113, SEQ ID NO: 114, SEQ ID NO: 115, SEQ ID NO: 116, SEQ ID NO: 117, SEQ ID NO: 118, SEQ ID NO: 119, SEQ ID NO: 120, SEQ ID NO: 121, SEQ ID NO: 122, SEQ ID NO: 123, SEQ ID NO: 124, SEQ ID NO: 125, SEQ ID NO: 126, SEQ ID NO: 127, SEQ ID NO: 128, SEQ ID NO: 129, SEQ ID NO: 130, SEQ ID NO: 131, SEQ ID NO: 132, SEQ ID NO: 133, SEQ ID NO: 134, SEQ ID NO: 135, SEQ ID NO: 136, SEQ ID NO: 137, SEQ ID NO: 138, SEQ ID NO: 139, SEQ ID NO: 140, SEQ ID NO: 141, SEQ ID NO: 142, SEQ ID NO: 143, SEQ ID NO: 144, SEQ ID NO: 145, SEQ ID NO: 146, SEQ ID NO: 147, SEQ ID NO: 148, SEQ ID NO: 149, SEQ ID NO: 150, SEQ ID NO: 151, SEQ ID NO: 152, SEQ ID NO: 192, SEQ ID NO: 193, SEQ ID NO: 194 and SEQ ID NO: 195.

It is understood herein that the low repetitive nucleotide sequences or the units or modules or building blocks provided herein can be permuted or combined with each other in any combination as long as the assembled nucleotide sequence comprises a low repetitive nucleotide sequence according to the present invention. Exemplary assembled nucleotide sequences are presented herein below, e.g., SEQ ID NO: 38 or herein depicted as PAS#1f/1c/1b(600), SEQ ID NO: 39 or herein depicted as PAS#1d/1f/1c/1b(800), SEQ ID NO: 40 or herein depicted as PAS#1h/1e/1i(600), SEQ ID NO: 41 or herein depicted as PAS#1j/1h/1e/1i(800), SEQ ID NO: 42 herein depicted as PA#1d/1c/1b(600), SEQ ID NO: 43 or herein depicted as PA#1i/1h/1g/1f(800), SEQ ID NO: 44 or herein depicted as PA#1e/1d/1c/1b(800), SEQ ID NO: 45 or herein depicted as PA#1i/1h/1g/1f/1e/1d/1c/1b(1600), SEQ ID NO: 153, SEQ ID NO: 154, SEQ ID NO: 155, SEQ ID NO: 156, SEQ ID NO: 157, SEQ ID NO: 158, SEQ ID NO: 159, SEQ ID NO: 160, SEQ ID NO: 161, SEQ ID NO: 162, SEQ ID NO: 163, SEQ ID NO: 164, SEQ ID NO: 165, SEQ ID NO: 166, SEQ ID NO: 167, SEQ ID NO: 168, SEQ ID NO: 169, SEQ ID NO: 170, SEQ ID NO: 171, SEQ ID NO: 172, SEQ ID NO: 173, SEQ ID NO: 174, SEQ ID NO: 175, SEQ ID NO: 176, SEQ ID NO: 177, SEQ ID NO: 178, SEQ ID NO: 179, SEQ ID NO: 180, SEQ ID NO: 181, SEQ ID NO: 182, SEQ ID NO: 184, SEQ ID NO: 185, SEQ ID NO: 186, SEQ ID NO: 187, SEQ ID NO: 188, SEQ ID NO: 189, SEQ ID NO: 190, and/or SEQ ID NO: 191.

Therefore, the nucleic acid molecule provided herein has, comprises or is said assembled nucleotide sequence in preferred aspects of the invention.

As documented in the appended examples (see Example 1), a long nucleotide sequence according to the invention may be assembled in a step-wise manner. The assembled nucleotide sequence maintains the low repetitiveness. It is demonstrated in the appended examples how an

- 43 -

exemplary nucleotide sequence as given in SEQ ID NO: 39 or herein depicted as PAS#1d/1f/1c/1b(800) comprising 2400 nucleotides and encoding 800 amino acids is assembled.

It is understood herein that these assembled nucleotide sequences are low repetitive nucleotide sequences. For example, the appended examples document that the exemplary nucleotide sequence as given in SEQ ID NO: 39 or depicted herein as PAS#1d/1f/1c/1b(800) sequence shows no repeats in case of a repeat window of 15, or only one single 14 nucleotide repeat, within the entire nucleotide sequence of 2400 nucleotides; see Figure 2B. For comparison, the long nucleotide sequence as disclosed in the prior art comprises repetitive nucleotide sequences as demonstrated exemplarily in case of PAS#1a(600) given herein as SEQ ID NO: 12. Accordingly, the long nucleic acid molecules according to the present invention have low repetitive nucleotide sequences and thus overcome technical challenges associated with repeated nucleotide stretches as mentioned above.

Nucleic acid molecules and related nucleic acid molecules (like variants, fragments, nucleic acid molecules having an identity of at least 66 %, e.g. at least 66.6% to the specific nucleotide sequences encoding a polypeptide consisting of proline and alanine; or like variants, fragments, nucleic acid molecules having an identity of at least 56 %, e.g. at least 56.6% to the specific nucleotide sequences encoding a polypeptide consisting of proline, alanine and serine as provided and defined herein, and the like) comprise or are low repetitive nucleotide sequences encoding the polypeptide, which forms the random coil conformation increasing the in vivo/vitro stability.

Nucleic acid molecules and related nucleic acid molecules have, comprise or are low repetitive nucleotide sequences encoding polypeptides that form random coil conformation and increase the in vivo/vitro stability of a biologically or pharmacologically active protein or drug. Said related nucleic acid molecules comprise or are variants and fragments of said nucleic acid molecules. Said related nucleic acid molecules have an identity of at least 66%, e.g. at least 66.6%, to a specific nucleotide sequence encoding a polypeptide consisting of proline and alanine or having an identity of at least 56%, e.g. at least 56.6%, to a specific nucleotide sequence encoding a polypeptide consisting of proline, alanine and serine as provided and defined herein, and the like.

In certain aspects, the nucleic acid molecule comprising an orthologous/homologous/identical/similar (and thus related) nucleotide sequence encoding a polypeptide consisting of proline and alanine is at least 66 %, e.g. at least 66.6% homologous/identical to the nucleotide sequence as, inter alia, shown in SEQ ID NOs: 28, 29,

- 44 -

30, 31, 32, 33, 34, 35, 36, 37, 42, 43, 44, 45, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 192 and 193. More preferably, the nucleic acid molecule comprising the orthologous/homologous/identical/similar (and thus related) nucleotide sequence encoding a polypeptide consisting of proline and alanine is at least 68%, 70%, 75%, 80%, 85%, 90%, 92%, 93%, 94%, 95%, 96%, 97% or 98% homologous/identical to the nucleotide sequence as, inter alia, shown in 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 42, 43, 44, 45, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 192 and 193, wherein the higher values are preferred. Most preferably, the nucleic acid molecule comprising the orthologous/homologous/identical/similar (and thus related) nucleotide sequence encoding a polypeptide consisting of proline and alanine is at least 99% homologous/identical/similar to the nucleotide sequence as, inter alia, shown in 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 42, 43, 44, 45, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 192 and 193.

In further aspects, the nucleic acid molecule comprising an orthologous/homologous/identical/similar (and thus related) nucleotide sequence encoding a polypeptide consisting of proline, alanine and serine is at least 56 %, e.g at least 56.6% homologous/identical/similar to the nucleotide sequence as, inter alia, shown in SEQ ID NOs: 19, 20, 21, 22, 23, 24, 25, 26, 27, 38, 39, 40, 41, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 174, 175, 176, 177, 178, 179, 180, 181, 182, 184, 185, 186, 187, 188, 189, 190, 191, 194 and 195. More preferably, the nucleic acid molecule comprising the orthologous/homologous/identical/similar (and thus related) nucleotide sequence encoding a polypeptide consisting of proline, alanine and serine is at least 58%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97% or 98% homologous/identical to the nucleotide sequence as, inter alia, shown in 19, 20, 21, 22, 23, 24, 25, 26, 27, 38 39, 40, 41, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 174, 175, 176, 177, 178, 179, 180, 181, 182, 184, 185, 186, 187, 188, 189, 190, 191, 194 and 195. Most preferably, the nucleic acid molecule comprising the orthologous/homologous/identical/similar (and thus related) nucleotide sequence encoding a polypeptide consisting proline, alanine and serine is at least 99% homologous/identical/similar to the nucleotide sequence as, inter alia, shown in SEQ ID

- 45 -

NOs: 19, 20, 21, 22, 23, 24, 25, 26, 27, 38, 39, 40, 41, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 174, 175, 176, 177, 178, 179, 180, 181, 182, 184, 185, 186, 187, 188, 189, 190, 191, 194 and 195. The above defined related nucleotide sequences can also be encompassed in longer or shorter isoforms, spliced variants or fusion constructs.

In certain aspects, the nucleic acid molecule provided herein may hybridize under stringent conditions to the complementary strand of the nucleotide sequence as, inter alia, shown in SEQ ID NOs: 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194 and 195. The term "hybridization" or "hybridizes" as used herein may relate to hybridizations under stringent or non-stringent conditions. If not further specified, the conditions are preferably stringent. Said hybridization conditions may be established according to conventional protocols described, e.g., in Sambrook (2001) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, 3rd ed., New York; Ausubel (1989) *Current Protocols in Molecular Biology*, Green Publishing Associates and Wiley Interscience, New York, or Higgins (1985) *Nucleic acid hybridization, a practical approach*, IRL Press Oxford, Washington DC. The setting of suitable conditions is well within the skill of the artisan and can be determined according to protocols described in the art. Thus, the detection of only specifically hybridizing sequences will usually require stringent hybridization and washing conditions such as, for example, 0.1 x saline sodium citrate buffer (SSC), 0.1% w/v SDS at 65°C or 2 x SSC, 60°C, 0.1 % w/v SDS. Low stringent hybridization conditions for the detection of homologous or not exactly complementary sequences may, for example, be set at 6 x SSC, 1% w/v SDS at 65°C. As is well known, the length of the nucleic acid probe and the composition of the nucleic acid to be determined constitute further parameters of the hybridization conditions.

In accordance with the present invention, the term "homology" or "percent homology" or "identical" or "percent identity" or "percentage identity" or "sequence identity" in the context of two or more nucleotide sequences refers to two or more sequences or subsequences that are the same, or that have a specified percentage of nucleotides that are the same (preferably at least 66 %, e.g., at least 66.6% identity in case of the nucleic acid molecule encoding the polypeptide consisting of proline and alanine, more preferably at least 68%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97% or 98% identity, most preferably at least

- 46 -

99% identity; or preferably at least 56 %, e.g., at least 56.6% identity in case of the nucleic acid molecule encoding proline, alanine and serine, more preferably at least 58%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97% or 98% identity, most preferably at least 99% identity) when compared and aligned for maximum correspondence over a window of comparison (preferably over the full length), or over a designated region as measured using a sequence comparison algorithm as known in the art, or by manual alignment and visual inspection.

Sequences having, for example, 70% to 90% or greater sequence identity may be considered to be substantially identical or similar. Such a definition also applies to the complement of a test nucleic acid sequence. Preferably, the described identity exists over a region that is at least about 15 to 25 nucleotides in length, more preferably, over a region that is at least about 50 to 100 nucleotides in length, more preferably over a region that is at least about 125 to 200 nucleotides in length, more preferably over a region that is at least about 225 to 300 nucleotides in length, even more preferably over a region that is at least about 325 to 600 nucleotides in length, even more preferably over a region that is at least about 625 to 800 nucleotides in length, and most preferably, over a region that is at least about 825 to 1200 nucleotides in length. Those having skill in the art will know how to determine percent identity between/among sequences using, for example, algorithms such as those based on CLUSTALW computer program (Thompson (1994) Nucl. Acids Res. 2:4673-4680), CLUSTAL Omega (Sievers (2014) Curr. Protoc. Bioinformatics 48:3.13.1-3.13.16) or FASTDB (Brutlag (1990) Comp App Biosci 6: 237-245). Also available to those having skill in this art are the BLAST, which stands for Basic Local Alignment Search Tool, and BLAST 2.0 algorithms (Altschul, (1997) Nucl. Acids Res. 25:3389-3402; Altschul (1990) J. Mol. Biol. 215:403-410). The BLASTN program for nucleic acid sequences uses as defaults a word length (W) of 11, an expectation (E) of 10, M=5, N=4, and a comparison of both strands. The BLOSUM62 scoring matrix (Henikoff (1992) Proc. Natl. Acad. Sci. U.S.A. 89:10915-10919) uses alignments (B) of 50, expectation (E) of 10, M=5, N=4, and a comparison of both strands.

In order to determine whether a nucleotide residue in a given nucleotide sequence corresponds to a certain position in the nucleotide sequence of, e.g., SEQ ID NOs: 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194 and 195, respectively, the skilled person can use means and methods

- 47 -

well known in the art, e.g., alignments, either manually or by using computer programs such as those mentioned herein. For example, BLAST 2.0 can be used to search for local sequence alignments. BLAST or BLAST 2.0, as discussed above, produces alignments of nucleotide sequences to determine sequence similarity. Because of the local nature of the alignments, BLAST or BLAST 2.0 is especially useful in determining exact matches or in identifying similar or identical sequences.

As shown in the appended examples, means and methods are provided herein to quickly and easily clone the long low repetitive nucleotide sequences or long low repetitive nucleic acid molecules comprising said nucleotide sequences; see, e.g., Examples 1, 2, 5, 7 and 10. As demonstrated in the appended examples, exemplary vectors are provided herein that are particularly suitable to assemble nucleotide sequences of the invention into longer constructs. As mentioned above, the nucleic acid molecules or nucleotide sequences provided herein can be assembled in a step-wise manner to build up long nucleic acid molecules or long nucleotide sequences. It is envisaged herein that the size of the assembled nucleic acid molecule or nucleotide sequence is not limited or is limited by practical circumstances, such as plasmid size or transformation/transfection efficiency, at best.

An exemplary assembly of a nucleic acid molecule comprising a low repetitive nucleotide sequence or a nucleic acid molecule encoding a repetitive amino acid sequence of proline, alanine and serine of a length of about 2400 nucleotides (encoding a polypeptide with a length of 800 amino acid residues) is demonstrated in the appended Example 2. It is shown therein that first the nucleotide sequence unit (building block) PAS#1b(200) as given in SEQ ID NO: 19, then the nucleotide sequence unit PAS#1c(200) as given in SEQ ID NO: 20, and subsequently the nucleotide sequence unit PAS#1f(200) as given in SEQ ID NO: 23, and finally the nucleotide sequence unit PAS#1d(200) as given in SEQ ID NO: 21 is cloned and assembled in the exemplary pXL2 vector. This exemplary assembled nucleic acid molecule or assembled nucleotide sequence is herein depicted as PAS#1d/1f/1c/1b(800) and is given in SEQ ID NO: 39. In addition, it is proven herein that these assembled inventive nucleic acid molecules have or comprise low repetitive nucleotide sequences. In particular, it is shown in Example 2 that this exemplary assembled nucleotide sequence as given in SEQ ID NO: 39 possesses only repeat sequences of a maximum length of 14 nucleotides.

The invention also relates to a method for providing the inventive nucleic acid molecules; see, for example, Figure 1E, which illustrates an exemplary procedure to assemble longer nucleotide sequences. Further, the invention relates to a vector that is particularly suitable to assemble the low repetitive nucleic acid molecules encoding the PA-rich polypeptides.

- 48 -

In accordance with the above, the nucleic acid molecule provided herein can encode a polypeptide consisting of proline and alanine.

Accordingly, the nucleic acid molecule provided herein can be selected from the group consisting of:

- (a) the nucleic acid molecule comprising at least one nucleotide sequence selected from the group consisting of SEQ ID NO: 28, SEQ ID NO: 29, SEQ ID NO: 30, SEQ ID NO: 31, SEQ ID NO: 32, SEQ ID NO: 33, SEQ ID NO: 34, SEQ ID NO: 35, SEQ ID NO: 36, SEQ ID NO: 37, SEQ ID NO: 87, SEQ ID NO: 88, SEQ ID NO: 89, SEQ ID NO: 90, SEQ ID NO: 91, SEQ ID NO: 92, SEQ ID NO: 93, SEQ ID NO: 94, SEQ ID NO: 95, SEQ ID NO: 96, SEQ ID NO: 97, SEQ ID NO: 98, SEQ ID NO: 99, SEQ ID NO: 100, SEQ ID NO: 101, SEQ ID NO: 102, SEQ ID NO: 103, SEQ ID NO: 104, SEQ ID NO: 105, SEQ ID NO: 106, SEQ ID NO: 107, SEQ ID NO: 108, SEQ ID NO: 109, SEQ ID NO: 110, SEQ ID NO: 111, SEQ ID NO: 112, SEQ ID NO: 113, SEQ ID NO: 114, SEQ ID NO: 115, SEQ ID NO: 116, SEQ ID NO: 117, SEQ ID NO: 118, SEQ ID NO: 119, SEQ ID NO: 120, SEQ ID NO: 121, SEQ ID NO: 122, SEQ ID NO: 192 and SEQ ID NO: 193;
- (b) the nucleic acid molecule comprising the nucleotide sequence consisting of SEQ ID NO: 42, SEQ ID NO: 43, SEQ ID NO: 44, SEQ ID NO: 45, SEQ ID NO: 153, SEQ ID NO: 154, SEQ ID NO: 155, SEQ ID NO: 156, SEQ ID NO: 157, SEQ ID NO: 158, SEQ ID NO: 159, SEQ ID NO: 160, SEQ ID NO: 161, SEQ ID NO: 162, SEQ ID NO: 163, SEQ ID NO: 164, SEQ ID NO: 165, SEQ ID NO: 166, SEQ ID NO: 167, SEQ ID NO: 168, SEQ ID NO: 169, SEQ ID NO: 170, SEQ ID NO: 171, SEQ ID NO: 172, and/or SEQ ID NO: 173;
- (c) the nucleic acid molecule hybridizing under stringent conditions to the complementary strand of the nucleotide sequence as defined in (a) or (b);
- (d) the nucleic acid molecule comprising the nucleotide sequence having at least 66.7% identity to the nucleotide sequence as defined in any one of (a), (b) and (c); and
- (e) the nucleic acid molecule being degenerate as a result of the genetic code to the nucleotide sequence as defined in (a) or (b).

In accordance with the above, the nucleic acid molecule provided herein can encode a polypeptide consisting of proline, alanine and serine.

Accordingly, the nucleic acid molecule provided herein can be selected from the group consisting of:

- 49 -

- (a) the nucleic acid molecule comprising at least one nucleotide sequence selected from the group consisting of SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 26, SEQ ID NO: 27, SEQ ID NO: 123, SEQ ID NO: 124, SEQ ID NO: 125, SEQ ID NO: 126, SEQ ID NO: 127, SEQ ID NO: 128, SEQ ID NO: 129, SEQ ID NO: 130, SEQ ID NO: 131, SEQ ID NO: 132, SEQ ID NO: 133, SEQ ID NO: 134, SEQ ID NO: 135, SEQ ID NO: 136, SEQ ID NO: 137, SEQ ID NO: 138, SEQ ID NO: 139, SEQ ID NO: 140, SEQ ID NO: 141, SEQ ID NO: 142, SEQ ID NO: 143, SEQ ID NO: 144, SEQ ID NO: 145, SEQ ID NO: 146, SEQ ID NO: 147, SEQ ID NO: 148, SEQ ID NO: 149, SEQ ID NO: 150, SEQ ID NO: 151, SEQ ID NO: 152; SEQ ID NO: 194 and SEQ ID NO: 195;
- (b) the nucleic acid molecule comprising the nucleotide sequence selected from the group consisting of SEQ ID NO: 38, SEQ ID NO: 39, SEQ ID NO: 40, SEQ ID NO: 41, SEQ ID NO: 174, SEQ ID NO: 175, SEQ ID NO: 176, SEQ ID NO: 177, SEQ ID NO: 178, SEQ ID NO: 179, SEQ ID NO: 180, SEQ ID NO: 181, SEQ ID NO: 182, SEQ ID NO: 184, SEQ ID NO: 185, SEQ ID NO: 186, SEQ ID NO: 187, SEQ ID NO: 188, SEQ ID NO: 189, SEQ ID NO: 190, and SEQ ID NO: 191;
- (c) the nucleic acid molecule hybridizing under stringent conditions to the complementary strand of the nucleotide sequence as defined in (a) or (b);
- (d) the nucleic acid molecule comprising the nucleotide sequence having at least 56% identity to the nucleotide sequence as defined in any one of (a), (b) and (c);
- (e) the nucleic acid molecule being degenerate as a result of the genetic code to the nucleotide sequence as defined in (a) or (b).

In certain aspects, the invention relates to a method for preparing a nucleic acid molecule comprising the low repetitive nucleotide sequence of the invention, wherein the method comprises:

- (a) providing a vector comprising an upstream recognition sequence that is recognized by an endonuclease restriction enzyme and a downstream recognition sequence that is recognized by an endonuclease restriction enzyme,
wherein optionally said endonuclease restriction enzyme recognizing said downstream recognition sequence is different from the endonuclease restriction enzyme recognizing the upstream recognition sequence,
wherein said upstream recognition sequence and said downstream recognition sequence are in a reverse complementary orientation,
wherein said upstream recognition sequence comprises two recognition sequences for two different restriction enzymes,
wherein said downstream recognition sequence is comprised in the upstream recognition sequence, and/or

- 50 -

wherein said upstream recognition sequence and/or said downstream recognition sequence are recognition sites for restriction enzymes cleaving outside of the recognition sequence;

- (b) cleaving said vector of (a) with the restriction enzyme(s) recognizing said upstream and/or said downstream recognition sequence;
- (c) optionally, dephosphorylating said vector of (b) to prevent hybridization of the complementary sticky ends;
- (d) providing a nucleic acid molecule encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein the ends of the nucleotide sequence hybridize to the cleaved ends of the vector of (b) or (c); and
- (e) inserting said nucleic acid molecule of (d) into said cleaved vector in the presence of a DNA ligase.

In certain aspects of the invention, the method provided herein can be used for preparing longer nucleic acid molecules, wherein the nucleic acid molecule comprises an assembled low repetitive nucleotide sequence, wherein the method for preparing the nucleic acid molecule as described above additionally comprises:

- (f) cleaving said vector of (e) with a restriction enzyme recognizing either said upstream or said downstream recognition sequence;
- (g) optionally, dephosphorylating said vector of (f) to prevent hybridization of the complementary sticky ends;
- (h) providing a nucleic acid molecule encoding a polypeptide consisting of proline, alanine and, optionally, serine, wherein the ends of the nucleotide sequence hybridize to the cleaved ends of the vector of (f) or (g);
- (i) inserting said nucleotide sequence of (h) into said cleaved vector in the presence of a DNA ligase,
- (j) repeating steps (f) to (i) until the desired length of the nucleotide sequence is reached.

The person skilled in the art understands that the steps of this method can be iteratively repeated until the desired length of the assembled nucleotide sequence or the assembled nucleic acid molecule encoding the PA-rich polypeptide is obtained.

In order to provide means and methods to advantageously clone the nucleic acid molecules provided herein, a suitable cloning vector comprises an upstream recognition sequence that is recognized by an endonuclease restriction enzyme and a downstream recognition sequence that is recognized by an endonuclease restriction enzyme, wherein optionally said endonuclease restriction enzyme recognizing said downstream recognition sequence is different from said endonuclease restriction enzyme recognizing said upstream recognition sequence, wherein said

- 51 -

upstream recognition sequence and said downstream recognition sequence are in a reverse complementary orientation. Further, said upstream recognition sequence can comprise two recognition sequences for two different restriction enzymes. Further, said downstream recognition sequence can be comprised in the upstream recognition sequence. Thus, said downstream recognition sequence can be comprised in one of the upstream recognition sequences. The appended illustrative examples show the exemplary vectors and nucleic acids molecules as well as methods to provide those, in particular, Figures 1, 4, 6 and 8.

It is documented in the appended examples that it is particularly beneficial that the downstream recognition sequence is comprised in the upstream recognition sequence. By using such a strategy, the upstream recognition site may be employed to assemble further nucleotide sequences into the vector provided herein. Of course, the positions of the upstream and downstream restriction sites on such a vector are interchangeable. An exemplary cloning region with an inserted inventive nucleotide sequence is shown in Figure 1D. In this instance, the restriction enzyme recognizing the downstream recognition sequence also recognizes and cleaves the upstream recognition sequence. Therefore, the inventive nucleotide sequence or the nucleic acid molecule encoding the polypeptide consisting of proline, alanine and, optionally, serine may be excised from the vector provided herein by employing the restriction enzyme recognizing the downstream and upstream recognition sequence, thus allowing use for ligation with other nucleic acids or vectors, e.g. to allow gene expression or to create a coding region for a fusion protein. An example for this cloning strategy to achieve assembly of a long low repetitive nucleotide sequence is illustrated herein below; see e.g. Figure 1E.

In certain aspects, the first overhang may comprise a nucleotide triplet/codon encoding alanine, in particular GCC, but also GCT, GCA or GCG. In a preferred aspect, the nucleic acid molecule of the invention comprises two complementary 5'-overhangs, i.e., one 5'-overhang is on the coding strand and the other 5'-overhang is on the non-coding strand. In a particularly preferred embodiment, the 5'-overhang on the coding strand is GCC, and the 5'-overhang on the non-coding strand is GGC.

In certain further aspects, the first overhang may also comprise nucleotide triplets/codons encoding proline or serine, e.g., CCT, CCC, CCA, CCG, TCT, TCC, TCA, TCG, AGT or AGC. However, the person skilled in the art knows that the method for preparing the nucleic acid molecule provided herein is not limited to overhangs but that the nucleic acid molecule or the nucleotide sequence can, for example, also be ligated via blunt ends. As used herein, the term "overhang" relates to an end-portion of a nucleic acid strand as part of the double-stranded DNA molecule with no attached complement, also known as a sticky end. As used herein, the term "blunt end" relates to an end-portion of a DNA strand with no overhang. It is envisaged

- 52 -

that the length of the overhang is not limited; however, an overhang comprising 1 to 10 nucleotides seems to be particularly suitable. In the appended examples, an overhang of 3 nucleotides encoding the amino acid alanine was employed. This kind of triplet overhang offers the advantage that it is directly compatible with the reading frame for amino acid translation of the nucleic acid molecule according to this invention.

Notably, the overhang introduces one additional triplet to the nucleotide sequence or the nucleic acid molecule encoding the polypeptide consisting of proline, alanine and, optionally, serine. Accordingly, the inventive cloning strategy introduces one additional amino acid, e.g., alanine. This additional amino acid or the corresponding triplet/codon can be considered as part of the polypeptide consisting of proline, alanine and, optionally, serine or the encoding nucleic acid, respectively. Consequently, as the skilled person in the art is aware, when employing the cloning strategy described above for a nucleotide sequence comprising, e.g., 300 nucleotides, the excised nucleic acid molecule comprises 303 nucleotides if also counting both three-nucleotide 5'-overhangs. It is envisaged herein that depending on the length of the overhangs employed herein even more triplets/codons or, accordingly, more amino acid residues can be introduced by this cloning procedure.

Furthermore, when cloned on a vector, or as part of a longer synthetic DNA fragment (e.g., a string), the nucleic acid molecule or the nucleotide sequence provided herein comprises said upstream recognition sequence and/or said downstream recognition sequence, wherein said upstream recognition sequence and/or said downstream recognition sequence are recognition sites for restriction enzymes that cleave outside of the recognition sequence. The restriction enzymes employed herein, e.g. those of the type IIS class preferably cleave outside of their recognition sequence to one side, which results in a cleaved nucleic acid molecule or nucleotide sequence that does not comprise the recognition site(s), depending on the proper orientation of the asymmetric recognition sequence.

Also, in the vector provided herein the cleavage with such restriction enzymes maintains the recognition sequences for the restriction enzymes. The cloning site of the exemplary vector pXL2 is shown in Figure 1C. A nucleic acid molecule or a nucleotide sequence excised from this vector lacks the recognition sites of the restriction enzymes used for its cloning or excision, which is particularly advantageous for the assembly of longer nucleic acid molecules or longer nucleotide sequences, e.g., according to the procedure described herein, for the cloning on an expression vector or for creating a coding region for a fusion protein.

In a further embodiment, the exemplary vector pXL1 is provided (SEQ ID NO: 55) (and shown in Figure 1B). In this case, the cloned/inserted nucleic acid molecule or nucleotide sequence

- 53 -

according to this invention is flanked by two *SapI* restriction sites. Hence, the nucleic acid molecule encoding the polypeptide consisting of proline, alanine and, optionally, serine can be simply excised from this vector by digest/cleavage with a single restriction enzyme, i.e. *SapI* (or also *EarI*), which is particularly advantageous for subsequent cloning on an expression vector and/or for creating a coding region for a fusion protein.

As it is shown in the appended examples, the nucleic acid molecules or the nucleotide sequences provided herein are assembled in the vector pXL2 (SEQ ID NO: 48) in an iterative manner; see, e.g., Example 2. Therefore, one or both of the recognition sites in the vector may be employed to insert a further nucleotide sequence into the vector (on one side of an insert present) or, alternatively, to excise the entire (assembled) nucleotide sequence or nucleic acid molecule from the vector.

As a further advantage, in the method for producing the inventive nucleic acid molecule provided herein, the absence of the recognition sites for the restriction enzymes within the nucleic acid molecule according to this invention (i) prevents that an assembled longer nucleotide sequence provided herein is cleaved internally, e.g., between the units/modules of the assembled nucleotide sequences and (ii) leads to an unbiased transition between the encoded amino acid sequences on neighboring units/modules of the assembled nucleotide sequences or between the encoded amino acid sequences on the nucleotide sequence of the invention and the nucleotide sequence for a biologically active protein. This strategy can be termed "traceless" or "seamless" assembly or cloning.

It is demonstrated in the appended examples that restriction enzymes of the type IIS can be employed; see Examples 1 and 2. Restriction enzymes of this class have recognition sites separate from their cleavage sites and some of them, for example *SapI* and *EarI*, cleave outside of their recognition sequence on one side. It is shown herein below that said upstream recognition sequence on pXL2 is recognized by *SapI* and *EarI* and said downstream recognition sequence is recognized by *EarI*. Thus, said upstream recognition sequence has the nucleotide sequence 5'-GCTCTTC-3' and said downstream recognition sequence has the nucleotide sequence 5'-CTCTTC-3'. The skilled person in the art knows that the restriction enzymes are not particularly limited herein as long as they fulfill the purpose or limitations defined herein.

As used herein, the term "endonuclease restriction enzyme" relates to an endonuclease enzyme cutting/cleaving/hydrolyzing oligo/polynucleotides at the phosphodiester bond connecting certain nucleotides within a nucleic acid molecule or nucleotide sequence, e.g., DNA, at or near a specific recognition sequence comprising a stretch of nucleotides. Thus, restriction enzymes

- 54 -

catalyze the hydrolysis of phosphodiester bonds within a nucleotide sequence or nucleic acid molecule. Restriction enzymes are commonly classified into three types which differ in their structure and whether they cut/cleave their substrate at their recognition sequence or at a separate site. To cut/cleave double-stranded DNA, restriction enzymes usually cleave two phosphodiester bonds, one in each sugar-phosphate backbone (i.e. each strand) of the DNA double helix.

As used herein, the term “recognition sequence” relates to a specific sequence of nucleotides, e.g., 4 to 8 specific base pairs in length, which are recognized by a restriction enzyme.

As used herein, the term “cleaving” means that the nucleic acid molecule and/or the vector is cut/digested/hydrolyzed with a restriction enzyme. As mentioned above, the restriction enzyme cleaves a phosphodiester bond within a polynucleotide chain.

As used herein, the term “inserting” refers to ligation of the nucleic acid molecule into the vector through the action of an enzyme. Thereby, the ends of the polynucleotides are joined together by the formation of phosphodiester bonds between the 3'-hydroxyl group at the terminus of one polynucleotide with the 5'-phosphoryl group of another. The nucleic acid molecule encoding the polypeptide consisting of proline, alanine and, optionally, serine or the nucleotide sequence has ends that can hybridize to the cleaved ends of the vector. In preferred aspects of the invention, such ends are overhangs that are able to hybridize to the respective overhangs of the vector cleaved.

For the insertion of the nucleic acid molecule into the vector, it is preferable to dephosphorylate the vector to avoid a high background of recircularized vector DNA with no insert. An exemplary enzyme for dephosphorylation may be calf-intestinal alkaline phosphatase (CIP or CIAP) or shrimp alkaline phosphatase which both remove the phosphate group from the 5' end of digested polynucleotides.

As used herein, the terms “upstream” and “downstream” both refer to a relative position on a nucleic acid molecule or nucleotide sequence. Each strand of the nucleic acid molecule or the nucleotide sequence has a 5' end and a 3' end, so named for the carbon atoms on the deoxyribose (or ribose) sugar. In general, upstream and downstream relate to the 5' to 3' direction of the coding strand in which RNA transcription takes place. When considering double-stranded DNA, upstream is toward the 5' end of the coding strand for the gene or reading frame in question and downstream is toward the 3' end. Due to the anti-parallel nature of double-stranded DNA, this means that the 3' end of the non-coding strand is upstream of the gene and its 5' end is downstream.

As used herein, the term "nucleic acid molecule" or "nucleotide sequence" is intended to include nucleic acid molecules such as DNA molecules and RNA molecules. It is herein understood that the term "nucleotide sequence" is equal to the term "nucleic acid sequence" and that these terms can be used interchangeably herein. Said nucleic acid molecule or said nucleotide sequence may be single-stranded or double-stranded, but preferably is double-stranded DNA. The skilled person in the art knows that double-stranded DNA actually comprises two different nucleic acid molecules, with largely complementary nucleotide sequences (neglecting sticky ends if present), which are non-covalently associated/hybridized to form a double strand.

In one aspect of the invention, the nucleotide sequence or the nucleic acid molecule encoding the polypeptide consisting of proline, alanine and, optionally, serine is operably linked in the same reading frame to a nucleic acid encoding a biologically or pharmacologically active protein. In preferred aspects of the invention, the nucleotide sequence or the nucleic acid molecule encoding the polypeptide consisting of proline, alanine and, optionally, serine is operably linked in the same reading frame to a nucleic acid encoding a biologically or pharmacologically active protein. Thus, the nucleic acid molecule encodes a heterologous drug conjugate comprising the polypeptide consisting of proline, alanine and, optionally, serine and the biologically or pharmacologically active protein. As used herein, heterologous means that the nucleic acid molecule encoding the polypeptide consisting of proline, alanine and, optionally, serine and the biologically or pharmacologically active protein is not found in nature.

As used herein, the term "operably linked" refers to a juxtaposition, wherein the components in question are in a relationship permitting them to both function in their intended manner.

The nucleotide sequence encoding the polypeptide consisting of proline, alanine and, optionally, serine can be conjugated to the nucleotide sequence encoding the biologically or pharmacologically active protein seamlessly, i.e., no linkers intersperse these two sequences. Alternatively, a linker or a spacer structure is comprised between the random coil polypeptide and the biologically or pharmacologically active protein. Thus, in certain aspects of the invention, a nucleotide sequence encoding an amino acid linker is inserted between the nucleotide sequence encoding the proline/alanine-rich polypeptide and the nucleotide sequence encoding the biologically or pharmacologically active protein. An exemplary linker can be a protease sensitive cleavage site, a serine/glycine-linker, an affinity tag such as the His₆-tag or the *Strep*-tag II, a signal peptide, retention peptide, a targeting peptide like a membrane translocation peptide or additional effector domains, e.g., antibody fragments for tumour

- 56 -

targeting associated with an anti-tumour toxin or an enzyme for prodrug activation etc. The polypeptide comprising a linker/spacer can have a plasma protease cleavage site that allows the controlled release of said biologically active protein. Linkers/spacers of different types or lengths may be identified without undue burden to obtain optimal biological activity of specific proteins.

Linkers/spacers can cause an immune response in the subject receiving the fusion protein that carries a linker or spacer structure. Therefore, in preferred aspects of the invention, the nucleotide sequence encoding the proline/alanine-rich polypeptide is conjugated to the nucleotide sequence encoding the biologically or pharmacologically active protein seamlessly. As used herein, "seamless" means that the nucleotide sequence encoding the random coil polypeptide is directly conjugated to the nucleotide sequence encoding the biologically or pharmacologically active protein. Thus, no additional nucleotides are introduced that encode amino acid residues other than proline, alanine or, optionally, serine. As demonstrated in the appended examples, a seamless cloning was accomplished by using overhangs that encode the amino acid residue alanine; see e.g. Example 7. Therefore, the present invention furthermore relates to a method for preparing the nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding proline, alanine and, optionally, serine and a nucleotide sequence encoding the biologically or pharmacologically active protein or peptide, and wherein said nucleotide sequence is conjugated seamlessly to said nucleotide sequence encoding the biologically or pharmacologically active protein. It is also demonstrated in the appended examples and herein above that the seamless cloning can be used to assemble longer nucleotide sequences encoding PA-rich polypeptides. Thus, if the seamless cloning method is applied, the resulting nucleotide sequence encoding the polypeptide consisting of proline, alanine and, optionally, serine is a low repetitive nucleotide sequence as described herein.

As used herein, the term "biological activity" describes the biological effect of a substance on living matter, for example an organism. Accordingly, the term "biologically active protein" or "biologically active peptide" as used herein relates to proteins or peptides that are capable of inducing a biological effect in living cells/organisms that are exposed to said protein or polypeptide or peptide. In the context of the invention, the nucleotide sequence or the nucleic acid molecule encoding the polypeptide consisting of proline, alanine and, optionally, serine is operably linked in the same reading frame to a heterologous nucleic acid encoding a biologically active protein.

It is herein understood, that the encoded biologically active protein or peptide is an "amino acid sequence having and/or mediating biological activity" or is an "amino acid sequence with biological activity" and/or is an amino acid sequence having and/or mediating a

- 57 -

pharmacological activity. Also comprised in the terms "biologically active protein", "amino acid sequence having and/or mediating biological activity" or "amino acid sequence with biological activity" and/or "amino acid sequence having and/or mediating a pharmacological activity" are any proteins or peptides of interest (and functional fragments thereof, such as antibody fragments, fragments comprising extracellular or intracellular domain(s) of a membrane receptor, truncated forms of a growth factor or cytokine and the like) for which prolongation of the half-life, either in vivo or in vitro, is beneficial. The skilled person in the art knows that the random coil conformation of the polypeptide consisting of proline, alanine and, optionally, serine mediates increased in vivo and/or in vitro stability to the biologically/pharmacologically active ("functional") protein(s) or peptide(s), in particular, an increased plasma half-life.

In one embodiment of this invention, the encoded amino acid sequence having and/or mediating biological activity in accordance with the present invention may be deduced from any "protein of interest", i.e., any protein of pharmaceutical or biological interest or any protein that is useful as a therapeutically effective protein. Accordingly, said biologically active protein or peptide may be a pharmacologically active or therapeutically effective protein or peptide. Pharmacologically active or therapeutically effective proteins or peptides are any proteins or peptides that have a desired pharmacological, pharmaceutical and/or physiological effect. The effect may be (i) prophylactic in terms of completely or partially preventing and/or ameliorating a disease/medical condition/disorder or symptom thereof; and/or may be (ii) therapeutic in terms of partially or completely inhibiting the disease/medical condition/disorder, i.e., arresting its development, or relieving the disease/medical condition/disorder, i.e., causing regression of the disease/medical condition/disorder. Most preferably, said biologically active protein is a therapeutically effective protein, e.g. for use as a vaccine. Thus, said biologically active protein can also be used in vaccination.

Furthermore, said biologically active protein can be a diagnostically relevant protein. As used herein, a "diagnostically relevant protein" relates to a protein or polypeptide that is employed in diagnosis. In the context of the present invention, diagnosis relates to the recognition and (early) detection of a disease, e.g. cancer and tumor, or a clinical condition in a subject. It may also comprise differential diagnosis. Also, the assessment of the severity of a disease or clinical condition may in certain embodiments be encompassed by the term "diagnosis". In particular, the polypeptides and/or drug conjugates encoded by the nucleic acid molecules of the present invention can be used for in vitro or in vivo diagnostics. For example, the encoded polypeptides and/or drug conjugates can be used in methods for medical imaging. The encoded PA-rich polypeptides and/or drug conjugates are in particular suitable for these methods as they have an improved uptake into tumor cells. Furthermore, the encoded PA-rich polypeptides

- 58 -

and/or drug conjugates show a higher contrast between tumor cells and blood or healthy cells/tissue.

The encoded biologically active protein is selected from the group consisting of a binding protein, an antibody fragment, a cytokine, a growth factor, a hormone, an enzyme, a protein vaccine, a peptide vaccine, or a peptide or a peptidomimetic. As used herein, a "peptide" preferably comprises/consists of up to 50 amino acid residues, whereas a "protein" preferably comprises/consists of 50 or more amino acid residues.

As used herein, the term "binding protein" relates to a molecule that is able to specifically interact with (a) potential binding partner(s) so that it is able to discriminate between said potential binding partner(s) and a plurality of molecules different from said potential binding partner(s) to such an extent that, from a pool of said plurality of different molecules as potential binding partner(s), only said potential binding partner(s) is/are bound, or is/are significantly bound. Methods for the measurement of binding activity between a binding protein and a potential binding partner are known in the art and can be routinely performed, e.g., by using enzyme-linked immunosorbent assay (ELISA), isothermal titration calorimetry (ITC), equilibrium dialysis, pull down assays, microscale thermophoresis, fluorescence titration or surface plasmon resonance (SPR) spectroscopy using, e.g., a Biacore instrument.

Exemplary binding proteins/binding molecules which are useful in the context of the present invention include, but are not limited to antibodies, antibody fragments such as Fab fragments, Fab' fragments, F(ab')₂ fragments, single chain variable fragments (scFv), (single) domain antibodies, in particular those derived from camelids, llamas or sharks, isolated variable regions of antibodies (VL and/or VH regions), in particular those from humans or primates, CDRs, immunoglobulin domains, CDR-derived peptidomimetics, lectins, fibronectin domains, tenascin domains, protein A domains, SH3 domains, ankyrin repeat domains, and lipocalins or various types of scaffold-derived binding proteins as described, for example, in Skerra (2000) *J. Mol. Recognit.* 13:167-187, Gebauer (2009) *Curr. Opin. Chem. Biol.* 13:245-255 or Binz (2005) *Nat. Biotechnol.* 23:1257-1268.

Other exemplary encoded biologically, pharmacologically active proteins or therapeutically effective proteins of interest that are useful in the context of the present invention include, but are not limited to, interleukin receptor antagonist, interleukin-1 receptor antagonist like EBI-005 or anakinra, leptin, acetylcholinesterase, activated protein C (drotrecogin), activin receptor IIB antagonist, adenosine deaminase, agalsidase alfa, agonist of toll-like receptor 5 like entolimod, alpha-1 antitrypsin, alpha-1 proteinase inhibitor, alpha-galactosidase, alpha-human atrial natriuretic peptide, alpha-N-acetylglucosaminidase, alteplase, amediopase, amylin, amylin

analogue, ANF-Rho, angiotensin (1-7), angiotensin II, angiotensin-converting-enzyme 2, anti-epithelial cell adhesion molecule single-chain antibody fragment, antithrombin alfa, antithrombin III, apoptosis inducing enzyme mi-APO, arginine deiminase, asparaginases like calaspargase, pegaspargase, crisantaspase, B domain deleted factor VIII like beroctocog alfa or octofactor, bectumomab (Lymphoscan), bile salt stimulated lipases like bucelipase alfa, binding protein directed against the respiratory syncytial virus like pavlizumab, bone morphogenetic proteins like BMP-2 (dibotermine alfa) or BMP-6, bouganin, bovine carboxyhemoglobin, bovine growth hormone, C1-Esterase-Inhibitor, C3 exoenzyme protein, carboxyhemoglobin, CD19 antagonist, CD20 antagonist like rituxan, CD3 receptor antagonist, CD40 antagonist, CD40L antagonist like dapirolizumab or Antova, cerebroside sulfatase, cethrin like VGX-210, chondroitin lyase, coagulation factor IX like nonacog gamma, conacog beta, albutrepenonacog alfa, coagulation factor VIIa like eptacog alfa, marzeptacog alfa, vatreptacog alfa, oreptacog alfa, coagulation factor VIII like susoctocog alfa, damoctocog alfa, turoctocog alfa, ruriococog alfa, efmoroctocog alfa, efraloctocog alfa, simoctocog alfa, coagulation factor X, coagulation factor XIII like catridecacog, collagenase of clostridium histolyticum, complement factor C3 inhibitor, complement receptor 5a antagonist, corticotrophin releasing factor, CSF1 receptor antagonists like FPA008, CSF1R antagonist, CTLA-4 antagonist like ipilimumab, cyanovirin-N, deoxyribonuclease I like dornase alfa, EGFR receptor antagonist, elastases like human type I pancreatic elastase like vonapanitase, endostatin, enkastim, epidermal growth factor, erythropoietin alfa, erythropoietin zeta, FcγIIB receptor antagonists, fibrinogenase, fibrinolytic enzyme like brinase, fibroblast growth factor 1 (human acidic fibroblast growth factor), fibroblast growth factor 18, fibroblast growth factor 2 (human basic fibroblast growth factor), fibroblast growth factor 21, fibroblast growth factor receptor 2 antagonists like FPA144, Fms-like tyrosine kinase 3 ligand, follicle-stimulating hormones like follitropin alfa or follitropin beta, fragment of human bactericidal/permeability-increasing protein 21 (opebacan/rBPI 21), gelonin, glucagon receptor agonist, glycoprotein IIb/IIIa antagonist like abciximab, glycosaminoglycan-degrading enzymes like condoliase, gp120/gp160, granulocyte colony stimulating factor (G-CSF), granulocyte macrophage colony stimulating factor (GM-CSF), heat-shock protein hsp 65 from Mycobacterium BCG fused with transcription factor E7 (verpasep caltespen), hepatocyte growth factor, hepatocyte growth factor receptor (HGFR) antagonist, hepcidin antagonist, Her2/neu receptor antagonist like herceptin, heterodimeric 15:IL-15Ra (hetIL-15), hirudin, hsp70 antagonist, human acid sphingomyelinase, human chorionic gonadotropin like choriogonadotropin alfa, human enzyme acid α-glucosidases like reveglucosidase alfa or alglucosidase alfa, human growth hormone, human keratinocyte growth factor (KGF), human matrix metalloproteinase, human myelin basic protein fragment, human osteogenic protein 1, human osteogenic protein-1, human parathyroid hormone, human thrombomodulin alfa, hyaluronidase like rHuPH20, hyaluronidases like human hyaluronidase PH-20 (vorhyaluronidase alfa), hyaluronidase or bovyhyaluronidase, hydrolytic lysosomal

- 60 -

glucocerebroside-specific enzymes like glucocerebrosidase, velaglycerase alfa or taliglycerase alfa, iduronate-2-sulfatase, IgE antagonists like omalizumab, ilroquois homeobox protein 2 (IRX-2), insulin, insulin analog, integrin $\alpha 4\beta 1$ antagonist, interferon tau, interferon-alpha, interferon-alpha antagonist, interferon-alpha superagonist, interferon-alpha-n3 (Alferon N Injection), interferon-beta, interferon-gamma, interferon-lambda, interleukin 2 fusion proteins like DAB(389)IL-2, interleukin-11 like oprelvekin, interleukin-12, interleukin-17 receptor antagonist, interleukin-18 binding protein, interleukin-2, interleukin-22, interleukin-4 like pitrakinra, interleukin-4 mutein, interleukin-6 receptor antagonist, interleukin-7, interleukin-22 receptor subunit alpha (IL-22ra) antagonist, irisin, islet neogenesis associated protein, kallidinogenase, lactoferrin, lactoferrin fragment, lanoteplase, lipase enzymes like burlulipase, rizolipase, epafipase or sebelipase alfa, luteinizing hormone, lutropin alpha, lymphocyte expansion molecule, lysostaphin, mammalian gastric lipase enzyme (merispace), mannosidases like velmanase alfa, melanocortin-4 receptor agonist, MEPE-derived 23-amino acid peptide, methionyl human stem cell factor (ancestim), microplasmin, N-acetylgalactosamine-6-sulfatase like elosulfase alfa, N-acetylglucosaminidase, nasaruplase beta, nerve growth factor, neuregulin-1, neurotoxin (e.g. a clostridial neurotoxin, like a *Clostridium botulinum* neurotoxin (such as *Clostridium botulinum* neurotoxin serotype A, B, C, D, E, F or G, particularly *Clostridium botulinum* neurotoxin serotype A), neutrophil gelatinase-associated lipocalin, ocriplasmin, Ornithodoros moubata complement inhibitor (OmCI/Coversin), osteoprotegerin, P128 (StaphTAME), pamiteplase, parathormone (PTH), PD-1 antagonist, PDGF antagonist, pentraxin-2 protein, phage lysin like HY133, phenylalanine ammonia lyase like valiasa, phosphatases like tissue-nonspecific alkaline phosphatase or asfotase alfa, plasminogen, plasminogen variant like V10153, platelet derived growth factor-BB, porcine growth hormone, prohibitin-targeting peptide 1, proinsulin, protein A, protein C like drotrecogin, protein binding fibroblast growth factor receptor ligands like FP-1039, recombinant tissue factor pathway inhibitor (tifacogin), relaxin, relaxin analog like serelaxin, reteplase, rhPDGF-BB, ribonuclease like onconase or amphinase, senrebotase, serine protease inhibitors like conestat alfa, sfericase, sialidase, soluble complement receptor type 1, soluble DCC (deleted in colorectal cancer) receptor, soluble TACI receptor (atacept), soluble tumor necrosis factor I receptor (sTNF-RI), soluble tumor necrosis factor II receptor (sTNF-RII), soluble VEGF receptor Flt-1, soluble, human Fc γ IIB receptor, staphylokinase, streptokinase, sulfamidase, T-cell receptor ligand, tenecteplase, thrombopoiesis-stimulating protein (AMG-531), thrombopoietin, thrombospondin-1, thyroid hormone, thyrotropin-releasing hormone (TRH) analog like taltirelin, tissue plasminogen activator, tissue-type plasminogen activator like pamiteplase, tripeptidyl peptidase I, tumor necrosis factor (TNF α), tumour necrosis factor α antagonist, uricase like rasburicase or pegadricase, urodilatin, urofollitropin, urokinase, uteroglobin, VEGF antagonist like ranbizumab or bevacizumab, VEGF/PDGF antagonist, VEGF/PDGF antagonist like a multi-VEGF/PDGF DARPin or a fusion protein, viscumin, von

Willebrand factors like vonicog alfa. Interleukin receptor antagonist, especially interleukin-1 receptor antagonists, like EBI-005 or anakinra, and leptin, especially human leptin, or a mutant human leptin (huLeptin(W100Q), a human leptin mutant with a tryptophan to glutamine substitution at position 100 in the mature polypeptide chain), are preferred herein. The amino acid sequence of human leptin is disclosed e.g. in UniProtKB accession code P41159. Mutant huLeptin(W100Q) is described by Ricci (2006) Mutational approach to improve physical stability of protein therapeutics susceptible to aggregation, In Murphy (ed.) Misbehaving proteins: protein (mis)folding, aggregation, and stability, Springer, 1st edition, New York.

Exemplary peptides and peptidomimetics include but are not limited to adrenocorticotrophic hormone (ACTH), afamelanotide, alarelin, alpha 4 integrin inhibitor, anti-HIV fusion inhibitor (like enfurvitide, V2o, SC34EK, SC35EK, IQN17 or IZN17), angiotensin II type 2 (AT2) receptor agonist (like LT2), anti-idiotypic p53 peptide, amylin, amylin analog, astressin, atosiban, bacterial peptide fragment with anticancer and anti HIV activity (like ATP-01), bicyclic peptide (like TG-758), bivalirudin, bradykinin antagonist (like icatibant), bremelanotide, B-type natriuretic peptide, calcitonin, carbetocin, carfilzomib, chrysalin, cilengitide, C-type natriuretic peptide, colostrinin, corticotrophin releasing factor (like Xerecept, coysnthropin), CNGRCG tumor homing peptide, ω -conotoxin peptide (like ziconotide), C-peptide, danegaptide, defensin, ecallantide, elcatonin, eledoisin, exendin-4, exendin-4 analog (like exendin 9-39), ezrin peptide 1, fragments from the human matrix extracellular phosphoglycoprotein (like AC-100), galanin, gastric inhibitory polypeptide (GIP), GIP analog, glatiramer, glucagon, glucagon analog, glucagon-like peptide 1 (GLP-1), GLP-1 analog (like lixisenatide, liraglutide or semiglutide), glucagon-like peptide 2 (GLP-2), GLP-2 analog (like teduglutide), gonadorelin, gonadotropin-releasing hormone agonist (like goserelin, buserelin, triptorelin, leuprolide, protirelin, lecirelin, fertirelin or desiorelin), gonadotropin-releasing hormone antagonist (like abarelix, cetrorelix, degarelix, ganirelix or teverelix), grehlin, grehlin analog (like AZP-531), growth hormone-releasing hormone, growth hormone-releasing hormone analog (like sermorelin or tesamorelin), hematide, hepcidin mimetic peptide, histrelin, indolicidin, indolicidin analog (like omiganan), IgE down-regulating peptide (like SC-01), INGAP peptide (exsulim), insulin-like growth factor 1, insulin-like growth factor 2, Kv1.3 ion channel antagonist (like cgtxA, cgtxE or cgtxF), lanreotide, lectin binding peptide (like sv6B, sv6D, svC2, svH1C, svH1D or svL4), lanthipeptide, larazotide, linacotide, lusupultide, melanocortin-4 receptor agonist (like AZD2820), MEPE-derived 23-amino acid peptide, mitochondrial-derived peptide (like MOTS-c, humanin, SHLP-6 or SHLP-2), mutant of the insulin-like growth factor binding protein-2 (like I-HBD1), Nav ion channel modulators (like GTx1-15 or VSTx3), octreotide, proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitory peptide, peptide fragment of azurin, Phylomer, peptide antagonist to the MHC Class II-associated invariant peptide (CLIP) (like VG1177), peptide derived from a heat shock

- 62 -

protein (like enkastim), pexiganan, plovamer, pramlintide, prohibitin-targeting peptide 1, pro-islet peptide, peptide tyrosine tyrosine (PYY 3-36), RGD peptide or peptidomimetic, ramoplanin, secretin, sinapultide, somatostatin, somatostatin analog (like pasireotide or CAP-232), specifically targeted antimicrobial peptide (STAMP) (like C16G2), receptor agonist of the bone morphogenetic protein (like THR-184 or THR-575), stresscopin, surfaxin, Tc99m apcitide, teriparatide (PTH 1-34), tetracosactide, thymosin alpha 1, TLR2 inhibitory peptide, TLR3 inhibitory peptide, TLR4 inhibitory peptide, thymosin B4, thymosin B15, vasoactive intestinal peptide, vasopressin, vasopressin analog like (desompressin, felypressin or terlypressin).

Exemplary biologically active proteins of interest that are useful in the context of vaccination include but are not limited to AE37 peptide, bombesin-gastrin-releasing peptide, carcinoembryonic antigen (CEA), the capsid protein open reading frame 2 (ORF2) protein of hepatitis E virus, cholera toxin B, clumping factor A of staphylococcus aureus, diphtheria toxin, diphtheria toxin mutant (like CRM 197), E. coli heat labile enterotoxin, exotoxin A of pseudomonas aeruginosa, F protein of measles virus, glycoprotein E of japanese encephalitis (JE) virus, GPC3-derived peptide, hepatitis A polyprotein, HER2-derived peptide GP2, herregulin, Her2neu peptid, hepatitis B virus surface antigen (HbSAg), human glutamic acid decarboxylase protein isoform 65 kDa (rhGAD65), influenza hemagglutinin antigens (HA), influenza neuraminidase (NA), L1 protein of human papilloma virus, li-Key/HER2/neu hybrid peptide, lipoprotein on outer surface of borrelia burgdorferi (OspA), major outer capsid protein of rota virus, mucin-1 (MUC-1) peptide, Norwalk virus (rNVP) capsid protein, Parvovirus B19 VLP, peptide derived from granulocyte-macrophage colony-stimulating factor, porcine circovirus 2 capsid (PCV2 ORF2) protein, protein C of tick-borne encephalitis virus, protein E of tick-borne encephalitis virus, protein E of yellow fever virus, protein E-1 of rubella virus, protein G of rhabdoviridae, protein H of measles virus, protein H of paramyxoviridae, protein NS of yellow fever virus, protein N of paramyxoviridae, prostate-specific antigen E2 protein of the swine fever virus, protein VP6 of rota virus, protein VP7 of rota virus, spike protein from the SARS virus (D3252), protein VP1 of polio virus, protein VP4 of polio virus, Ras oncoprotein, sperm derived peptides (like YLP12, P10G, A9D, mFA-12-19, SP56 and or mFA-1117-136), tetanus toxin, tuberculin, tumor-associated peptides (TUMAPs) (like IMA901, IMA910 or IMA950), and the like.

- 63 -

In one aspect, the present invention relates to a nucleic acid molecule as disclosed herein, for example, a nucleic acid molecule comprising a nucleotide sequence encoding a polypeptide consisting of proline, alanine and serine, wherein the nucleotide sequence of said nucleic acid has a length of at least 300 nucleotides; or

the present invention relates in one aspect to a nucleic acid molecule disclosed herein that is operably linked in the same reading frame to a nucleic acid encoding a biologically active protein, for example a nucleic acid molecule comprising a nucleotide sequence encoding a polypeptide consisting of proline, alanine and serine, wherein the nucleotide sequence of said nucleic acid has a length of at least 300 nucleotides, wherein the nucleic acid molecule is operably linked in the same reading frame to a nucleic acid encoding a biologically active protein;

wherein said nucleotide sequence is not

```
ATGGGCAGCAGCCATCATCATCACCATCATGGTAGCCTGGTTCCGCGTAGCTCTTCTGCA
AGTCCGGCAGCACCGGCACCGGCTTCACCAGCTGCACCAGCACCTAGCGCACCGGCAGCA
TCTCCAGCAGCCCCTGCACCGGCAAGCCCTGCAGCTCCAGCACCGTCAGCACAGCAGCA
AGCCCAGCTGCTCCTGCTCCAGCGAGCCCAGCAGCGCCAGCTCCTAGTGCCCCCTGCTGCC
TCTCCTGCTGCTCCGGCACCAGCAAGTCCTGCTGCGCCTGCACCGAGTGCTCCGGGCTGCT
AGTCCTGCCGCACCAGCTCCGGCTAGTCCAGCTGCTCCAGCCCCTTCAGCTCCGGCAGCT
TCCCCTGCAGCGCCTGCCCTGCCAGTCCAGCGGCTCCTGCACCTAGTGCGCCTGCAGCT
TCACCGGCTGCCCTGCGCCAGCTTCTCCTGCGGCTCCAGCTCCATCTGCCCCAGCCGCA
TCCCCAGCGGCACCAGCTCCAGCTTCTCCGGCAGCGCCAGCACCTTCTGCGCCTGCCGCA
TCTCCTGCAGCACAGCGCCTGCGAGTCCTGCAGCTCCTGCTCCTTCAGCCCCTGCGGCA
AGTCCAGCAGCACAGCCCCAGCAAGCCCAGCCGCACCAGCACCATCTGCCCTGCAGCA
CCATTGTGAACAAGCAGTTTAACTATAAGGACCCGGTGAACGGTGTGGATATCGCGTAT
ATCAAAATCCCGAATGCGGGCCAGATGCAACCAGTCAAGGCGTTCAAGATTCATAACAAG
ATTTGGGTTATTCCGGAACGTGATACCTTCACCAATCCGGAAGAAGGCGACTTAAACCCG
CCGCCAGAAGCCAAACAAGTGCCGGTGAGCTACTATGATAGCACGTATCTTAGCACCGAT
AATGAAAAAGACAATTACCTGAAGGGCGTGACCAAGTTGTTGAGCGCATCTACAGTACC
GACTTAGGCCGCATGTTGTTGACGAGCATCGTTCGCGGTATCCCGTTCTGGGGCGGCTCG
ACCATTGATACCGAGTTGAAAGTCATTGACACGAACTGTATCAATGTTATCCAACCGGAC
```

GGCAGTTATCGCAGCGAGGAGTTAAATTTGGTCATCATCGGTCCAAGCGCAGATATTATT
CAGTTCGAATGCAAGAGCTTCGGCCATGAGGTCTTGAATTTGACGCGCAACGGTTACGGC
AGCACCCAATACATCCGCTTTAGCCCGGATTTACCTTTGGCTTCGAGGAGAGCTTGGAG
GTGGACACCAACCCGCTGTTAGGTGCCGGCAAATTCGCAACCGACCCGGCAGTGACGTTG
GCGCACGAATTGATTCATGCGGGTCACCGCTTATACGGTATCGCGATCAATCCGAATCGC
GTCTTTAAAGTCAATACCAACGCGTACTACGAAATGAGCGGCTTAGAGGTTAGCTTTGAA
GAATTACGCACCTTCGGTGGCCACGACGCCAAGTTCATCGACAGCCTGCAGGAAAATGAG
TTCCGCTTGTAATAATTACAATAAATTCAAGGACATCGCGAGCACCTTAAATAAAGCAAAG
AGCATTGTGGGCACCACCGCAAGCTTGCAGTACATGAAGAACGTATTTAAGGAAAAATAT
TTGTTGTGCGGAGGATACCAGCGGGAAATTCAGCGTCGATAAGCTGAAATTCGACAAATTG
TATAAAATGCTGACCGAGATTTACACCGAGGATAACTTCGTCAAGTTTTTTAAGGTGTTA
AATCGTAAGACCTATTTAACTTTGATAAAGCGGTGTTTAAATTAATATCGTGCCGAAG
GTGAATTACACCATCTACGATGGTTTTCAATTTACGCAACACGAATCTGGCGGCGAATTTT
AATGGCCAAACACCGAAATTAACAACATGAACCTTTACGAAGTTAAAGAATTTACGCGGC
TTATTCGAATTCTACAAGTTATTATGCGTGCGCGGCATCATTACCAGCAAGGCAGGTGCG
GGCAAGTCCTTGGTTCCGCGTGCGAGCGCCGGCGCGCGCTCAATGATCTGTGTATT
AAAGTCAATAACTGGGACCTGTTCTTCAGCCCGAGCGAGGATAACTTTACCAACGACTTA
AACAAAGGCGAGGAGATCACGAGCGATACGAACATCGAGGCGGCGGAGGAAAATATTAGC
CTGGACCTCATTCAGCAGTACTATCTGACGTTCAATTTTGACAATGAGCCGGAGAACATC
AGCATTGAAAATCTCAGCAGCGACATCATCGGTGAGTTGGAACCTGATGCCGAACATTGAA
CGCTTTCCGAACGGCAAATAATGAACCTGGACAAGTATACCATGTTCCATTACTTACGC
GCACAGGAATTTGAGCACGGCAAGAGCCGCATTGCGCTGACCAATAGCGTTAACGAGGCC
TTGTTAAATCCGAGCCGTGTCTACACGTTCTTCAGCAGCGATTATGTCAAAAAAGTGAAC
AAGGCGACCGAAGCCGCGATGTTTTTGGGCTGGGTGAGCAATTGGTTTACGATTTTACC
GACGAAACCAGCGAGGTGAGCACGACCGACAAAATTGCAGATATCACCATCATCATTCGG
TACATCGGTCCGGCGCTCAATATCGGCAATATGTTATACAAGGACGACTTTGTGGGCGCG
CTGATCTTTAGCGGCGCGGTTATCTTATTAGAATTCATCCCGGAGATCGCAATCCCGGTC
TTGGGCACCTTTGCGTTGGTGAGCTATATCGCGAATAAAGTGCTCACGGTCCAAACCATC
GATAACGCGCTCAGCAAGCGTAATGAGAAATGGGACGAGGTTTATAAGTATATCGTGACC
AACTGGTTAGCAAAAGTCAATACGCAGATCGATCTCATCCGCAAAAAAATGAAAGAAGCC
TTGGAAAATCAAGCGGAGGCAACCAAGCCATCATTAATTACCAGTATAACCAATATACC
GAAGAAGAAAAAACAATATCAACTTCAATATCGATGATTTGAGCAGCAAACTGAACGAG
AGCATTAACAAAGCGATGATTAACATCAACAAGTTCTTGAATCAATGCAGCGTGAGCTAT
CTCATGAACAGCATGATCCCGTATGGCGTCAAACGCTTGGAAGATTTTGACGCCAGCCTG
AAAGATGCGCTCCTCAAGTATATTTATGACAACCGCGGCACCCTCATTTGGCCAGGTGGAC
CGCTTGAAGGATAAAGTGAACAATACGCTCAGCACGGATATCCCGTTCCAGCTGAGCAAG
TACGTCGACAACCAGCGCTTACTGAGCACCTTTACCGAGTATATCAAGAACATCATTAAT
ACCAGCATCCTCAACTTGCGCTATGAGAGCAATCACCTGATCGACCTCAGCCGCTACGCC
AGCAAGATCAACATCGGCAGCAAGGTCAATTTGACCCGATCGATAAGAATCAGATCCAA
TTGTTTAACTTGAAAGCAGCAAGATCGAGGTTATCTTGAAGAACGCGATTGTGTACAAC
AGCATGTACGAGAACTTTAGCACGAGCTTCTGGATTCTGATCCCGAAGTATTTCAATAGC
ATTAGCCTGAATAACGAATATACCATTATCAACTGCATGGAAAATAATAGCGGCTGGAAG
GTGAGCTTAAATTACGGCGAGATCATTTTGACCTTACAGGATACCCAAGAAATCAAACAG
CGCGTCGTCTTTAAGTATAGCCAGATGATCAACATCAGCGATTACATCAACCGCTGGATC
TTCGTGACCATCACCAATAATCGCTTGAATAATAGCAAGATTTACATCAATGGTCGCTTG
ATTGATCAAAAACCGATCAGCAATCTCGGTAATATCCATGCCAGCAATAACATCATGTTT
AAGTTAGACGGTTGCCGCGATACCCACCGCTATATCTGGATCAAGTATTTTAACTTATTT
GATAAGGAACTCAACGAAAAGGAAATTAAGACTTATATGACAATCAGAGCAATAGCGGC
ATCCTGAAGGATTTCTGGGGCGACTACCTGCAGTACGATAAGCCGTACTATATGTTGAAC
TTGTATGACCCGAACAAATATGTCGATGTGAACAATGTGGGTATTCTGTGGCTATATGTAC

- 65 -

TTAAAGGGCCCGCGTGGTAGCGTGATGACCACGAATATTTACTTAAACAGCAGCTTATAC
CGCGGCACGAAGTTTATTATCAAGAAGTATGCCAGCGGCAACAAGGACAATATCGTCCGC
AACACGACCGTGTGTATATTAACGTGGTGGTGAAGAATAAAGAGTACCGCTTGGCCACG
AATGCGAGCCAGGCGGGCGTGGAAAAAATCTTGAGCGCGTTGGAGATCCCGGACGTCGGC
AACCTCAGCCAGGTTGTGGTGATGAAGTCTAAAAACGACCAGGGCATCACGAACAAGTGC
AAAATGAATTTGCAAGATAACAACGGCAACGACATCGGCTTTATTGGTTTTTCACCAGTTC
AATAACATCGCCAAACTCGTGGCCAGCAATTGGTATAACCGCCAAATTGAACGCAGCAGC
CGCACGCTCGGCTGTAGCTGGGAGTTCATCCCGGTGGACGATGGCTGGGGCGAGCGCCCG
CTCGGAGATCTGGTGCCACGCGGTTCCGCGAATTCGAGCTCCGTCGACAAGCTTTGGAGC
CACCCGCAGTTCGAAAAATAA

(SEQ ID No. 196)

In one aspect, the present invention relates to a nucleic acid molecule as disclosed herein, for example, a nucleic acid molecule comprising a nucleotide sequence encoding a polypeptide consisting of proline, alanine and serine, wherein the nucleotide sequence of said nucleic acid has a length of at least 300 nucleotides; or

the present invention relates in one aspect to a nucleic acid molecule disclosed herein that is operably linked in the same reading frame to a nucleic acid encoding a biologically active protein, for example a nucleic acid molecule comprising a nucleotide sequence encoding a polypeptide consisting of proline, alanine and serine, wherein the nucleotide sequence of said nucleic acid has a length of at least 300 nucleotides, wherein the nucleic acid molecule is operably linked in the same reading frame to a nucleic acid encoding a biologically active protein;

wherein said nucleotide sequence is not

- 66 -

ATGGGTAGCAGCCATCATCATCACCATCATGGTAGCCTGGTTCCGCGTAGCTCTTCTGCA
AGTCCGGCAGCACCCGGCACC GGCTTCACCAGCTGCACCAGCACCTAGCGCACCCGGCAGCA
TCTCCAGCAGCCCCCTGCACCGGCAAGCCCTGCAGCTCCAGCACCGTCAGCACCAGCAGCA
AGCCCAGCTGCTCCTGCTCCAGCGAGCCCAGCAGCGCCAGCTCCTAGTGCCCCTGCTGCC
TCTCCTGCTGCTCCGGCACCAGCAAGTCCTGCTGCGCCTGCACCGAGTGCTCCGGCTGCT
AGTCCTGCCGCACCAGCTCCGGCTAGTCCAGCTGCTCCAGCCCCCTTCAGCCCCCTGCAGCA
CCATTTGTGAACAAGCAGTTTAACTATAAGGACCCGGTGAACGGTGTGGATATCGCGTAT
ATCAAAATCCCGAATGCGGGCCAGATGCAACCAGTCAAGGCGTTCAAGATTCATAACAAG
ATTTGGGTATTCCGGAACGTGATACCTTCACCAATCCGGAAGAAGGCGATTTAAATCCG
CCGCCAGAAGCCAAACAAGTGCCGGTGAGCTACTATGATAGCACGTATCTTAGCACCGAT
AATGAAAAAGACAATTACCTGAAGGGCGTGACCAAGTTGTTTCGAGCGCATCTACAGTACC
GACTTAGGCCGCATGTTGTTGACGAGCATCGTTCGCGGTATCCCGTTCTGGGGCGGCTCG
ACCATTGATACCGAGTTGAAAGTCATTGACACGAACTGTATCAATGTTATCCAACCGGAC
GGCAGTTATCGCAGCGAGGAGTTAAATTTGGTCATCATCGGTCCAAGCGCAGATATTATT
CAGTTCGAATGCAAGAGCTTCGGCCATGAGGTCTTGAATTTGACGCGCAACGGTTACGGC
AGCACCCAATACATCCGCTTTAGCCCGGATTTACCTTTGGCTTCGAGGAGAGCTTGAG
GTGGACACCAACCCGCTGTTAGGTGCCGGCAAATTCGCAACCGACCCGGCAGTGACGTTG
GCGCACGAATTGATTCATGCGGGTCACCGCTTATACGGTATCGCGATCAATCCGAATCGC
GTCTTTAAAGTCAATACCAACGCGTACTACGAAATGAGCGGCTTAGAGGTTAGCTTTGAA
GAATTACGCACCTTCGGTGGCCACGACGCCAAGTTTCATCGACAGCCTGCAGGAAAATGAG
TTCCGCTTGTTACTATTACAATAAATTCAGGACATCGCGAGCACCTTAAATAAAGCAAAG
AGCATTTGTGGGCACCACCGCAAGCTTGCAGTACATGAAGAACGTATTTAAGGAAAAATAT
TTGTTGTCGGAGGATACCAGCGGGAAATTCAGCGTCGATAAGCTGAAATTCGACAAATTG
TATAAAATGCTGACCGAGATTTACACCGAGGATAACTTCGTCAAGTTTTTTAAGGTGTTA
AATCGTAAGACCTATTTAACTTTGATAAAGCGGTGTTTTAAATTAATATCGTGCCGAAG
GTGAATTACACCATCTACGATGGTTTTCAATTTACGCAACACGAATCTGGCGGCGAATTTT
AATGGCCAAAACACCGAAATTAACAACATGAACTTTACGAAGTTAAAGAATTTACGGGC
TTATTCGAATTCCTACAAGTTATTATGCGTGCGCGGCATCATTACCAGCAAGGCAGGTGCG
GGCAAGTCCTTGGTTCCGCGTGGCAGCGCCGGCGCCGGCGCGCTCAATGATCTGTGTATT
AAAGTCAATAACTGGGACCTGTTCTTCAGCCCGAGCGAGGATAACTTTACCAACGACTTA
AACAAAGGCGAGGAGATCACGAGCGATACGAACATCGAGGCGGCGGAGGAAAATATTAGC
CTGGACCTCATTACAGCAGTACTATCTGACGTTCAATTTTGACAATGAGCCGGAGAACATC
AGCATTGAAAATCTCAGCAGCGACATCATCGGTCAGTTGGAAGTATGATGCCGAACATTGAA
CGCTTTCCGAACGGCAAAAAATATGAACTGGACAAGTATACCATGTTCCATTACTTACGC
GCACAGGAATTTGAGCACGGCAAGAGCCGCATTGCGCTGACCAATAGCGTTAACGAGGCC
TTGTTAAATCCGAGCCGTGTCTACACGTTCTTCAGCAGCGATTATGTCAAAAAGTGAAC
AAGGCGACCGAAGCCGCGATGTTTTTGGGCTGGGTCGAGCAATTGGTTTACGATTTTACC

- 67 -

GACGAAACCAGCGAGGTGAGCACGACCGACAAAATTGCAGATATCACCATCATCATTCGG
TACATCGGTCCGGCGCTCAATATCGGCAATATGTTATACAAGGACGACTTTGTGGGCGCG
CTGATCTTTAGCGGCGCGGTTATCTTATTAGAATTCATCCCGGAGATCGCAATCCCGGTC
TTGGGCACCTTTGCGTTGGTGAGCTATATCGCGAATAAAGTGCTCACGGTCCAAACCATC
GATAACGCGCTCAGCAAGCGTAATGAGAAATGGGACGAGGTTTATAAGTATATCGTGACC
AACTGGTTAGCAAAAGTCAATACGCAGATCGATCTCATCCGCAAAAAAATGAAAGAAGCC
TTGGAAAATCAAGCGGAGGCAACCAAGCCATCATTAATTACCAGTATAACCAATATACC
GAAGAAGAAAAAACAATATCAACTTCAATATCGATGATTTGAGCAGCAAACTGAACGAG
AGCATTAAACAAAGCGATGATTAACATCAACAAGTTCTTGAATCAATGCAGCGTGAGCTAT
CTCATGAACAGCATGATCCCGTATGGCGTCAAACGCTTGGAAGATTTTGACGCCAGCCTG
AAAGATGCGCTCCTCAAGTATATTTATGACAACCGCGGCACCCTCATTGGCCAGGTGGAC
CGCTTGAAGGATAAAGTGAACAATACGCTCAGCACGGATATCCCGTTCCAGCTGAGCAAG
TACGTCGACAACCAGCGCTTACTGAGCACCTTTACCGAGTATATCAAGAACATCATTAAT
ACCAGCATCCTCAACTTGCGCTATGAGAGCAATCACCTGATCGACCTCAGCCGCTACGCC
AGCAAGATCAACATCGGCAGCAAGGTCAATTTTCGACCCGATCGATAAGAATCAGATCCAA
TTGTTTAACCTGGAAAGCAGCAAGATCGAGGTTATCTTGAAGAACGCGATTGTGTACAAC
AGCATGTACGAGAACTTTAGCACGAGCTTCTGGATTCTGATCCCGAAGTATTTCAATAGC
ATTAGCCTGAATAACGAATATAACCATTATCAACTGCATGGAAAATAATAGCGGCTGGAAG
GTGAGCTTAAATTACGGCGAGATCATTTGGACCTTACAGGATACCCAAGAAATCAAACAG
CGCGTCGTCTTTAAGTATAGCCAGATGATCAACATCAGCGATTACATCAACCGCTGGATC
TTCGTGACCATCACCATAATCGCTTGAATAATAGCAAGATTTACATCAATGGTCGCTTG
ATTGATCAAAAACCGATCAGCAATCTCGGTAATATCCATGCCAGCAATAACATCATGTTT
AAGTTAGACGGTTGCCGCGATACCCACCGCTATATCTGGATCAAGTATTTTAACTTATTT
GATAAGGAACTCAACGAAAAGGAAATTAAAGACTTATATGACAATCAGAGCAATAGCGGC
ATCCTGAAGGATTTCTGGGGCGACTACCTGCAGTACGATAAGCCGTACTATATGTTGAAC
TTGTATGACCCGAACAAATATGTCGATGTGAACAATGTGGGTATTTCGTGGCTATATGTAC
TTAAAGGGCCCGCTGGTAGCGTGATGACCACGAATATTTACTTAAACAGCAGCTTATAC
CGCGGCACGAAGTTTATTATCAAGAAGTATGCCAGCGGCAACAAGGACAATATCGTCCGC
AACAACGACCGTGTGTATATTAACGTGGTGGTGAAGAATAAAGAGTACCGCTTGGCCACG
AATGCGAGCCAGGCGGGCGTGGAATAATCTTGAGCGCGTTGGAGATCCCGGACGTCGGC
AACCTCAGCCAGGTTGTGGTGATGAAGTCTAAAAACGACCAGGGCATCACGAACAAGTGC
AAAATGAATTTGCAAGATAACAACGGCAACGACATCGGCTTTATTGGTTTTACACAGTTC
AATAACATCGCCAAACTCGTGGCCAGCAATTGGTATAACCGCCAAATTGAACGCAGCAGC
CGCACGCTCGGCTGTAGCTGGGAGTTCATCCCGGTGGACGATGGCTGGGGCGAGCGCCCG
CTCGGAGATCTGGTGCCACGCGGTTCCGCGAATTCGAGCTCCGTGACAAGCTTTGGAGC
CACCCGCAGTTCGAAAAATAA

(SEQ ID No. 197)

In one aspect, the present invention relates to a nucleic acid molecule, wherein said nucleic acid molecule consists of a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine,
wherein the nucleotide sequence of said nucleic acid has a length of at least 300 nucleotides,
wherein said nucleotide sequence has a Nucleotide Repeat Score (NRS) lower than 50,000,
wherein said Nucleotide Repeat Score (NRS) is determined according to the formula:

- 68 -

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{i=1}^{k(n)} f_i(n)}}{N_{tot}},$$

wherein

N_{tot} is the length of said nucleotide sequence,

n is the length of a repeat within said nucleotide sequence, and

$f_i(n)$ is the frequency of said repeat of length n ,

wherein if there is more than one repeat of length n , $k(n)$ is the number of said different sequences of said repeat of length n , otherwise $k(n)$ is 1 for said repeat of length n .

Furthermore, the present invention also relates to a vector comprising the nucleic acid molecule or the nucleotide sequence of the invention. The vector can also be employed in order to provide a nucleic acid molecule comprising (i) a nucleotide sequence encoding a polypeptide consisting of alanine, proline and, optionally, serine and (ii) a nucleotide sequence encoding a biologically active protein; see, e.g., Fig. 4 and Example 7.

Such a vector can be employed as an expression vector in order to express fusion proteins comprising the encoded random coil polypeptides and biologically active proteins. Accordingly, the encoded fusion protein encompasses (i) a random coil polypeptide, which is encoded by the low repetitive nucleotide sequence, coupled to (ii) a biologically active protein. Preferably, the random coil polypeptide consists of alanine, proline and, optionally, serine. An exemplary vector is given in SEQ ID NO: 56. In the appended examples, an exemplary method is demonstrated in order to provide such a vector or nucleic acid molecule; see e.g. Fig. 6 and Example 7.

In this inventive method, the vector provided herein comprises the nucleotide sequence encoding the biologically active protein and, in a second step, the nucleotide sequence encoding the PA-rich polypeptide is introduced into the vector. In order to introduce the nucleotide sequence encoding the proline/alanine-rich sequence into the vector comprising the nucleotide sequence encoding the biologically active protein, overhangs can be employed that comprise at least one nucleotide triplet/codon encoding e.g. alanine, proline and/or optionally serine; see above. Accordingly, such a triplet or codon can encode an amino acid that is considered part of the proline/alanine-rich sequence, in particular alanine.

Consequently, the method and vector provided herein avoid the introduction of additional amino acid linkers that may be introduced by utilizing conventional restriction sites. Therefore, the means and methods provided herein allow a seamless cloning of the inventive nucleic acid

- 69 -

molecule comprising (i) the nucleotide sequence encoding the polypeptide consisting of alanine, proline and, optionally, serine and (ii) the nucleotide sequence encoding the biologically active protein.

Many suitable vectors are known to those skilled in molecular biology. The choice of a suitable vector depends on the function desired, including plasmids, cosmids, viruses, bacteriophages and other vectors used conventionally in genetic engineering.

Preferably, the vector is a plasmid, more preferably a plasmid based on the generic *E. coli* expression vector pASK37, pASK75 or pXL2.

Methods which are well known to those skilled in the art can be used to construct various plasmids; see, for example, the techniques described in Sambrook (2001) loc cit. and Ausubel (1989) loc. cit. Typical plasmid vectors include, e.g., pQE-12, the pUCseries of plasmids, pBluescript (Stratagene), the pET series of expression vectors (Novagen) or pCARTOPO (Invitrogen), lambda gt11, pJOE, the pBBR1-MCS series, pJB861, pBSMuL, pBC2, pUCPKS, pTACT1. Typical vectors compatible with expression in mammalian cells include E-027 pCAG Kosak-Cherry (L45a) vector system, pREP (Invitrogen), pCEP4 (Invitrogen), pMC1neo (Stratagene), pXT1 (Stratagene), pSG5 (Stratagene), EBO-pSV2neo, pBPV-1, pDBPVMMTneo, pRSVgpt, pRSVneo, pSV2-dhfr, pIZD35, Okayama-Berg cDNA expression vector pcDV1 (Pharmacia), pRc/CMV, pcDNA1, pcDNA3 (Invitrogen), pcDNA3.1, pSPORT1 (GIBCO BRL), pGEMHE (Promega), pLXIN, pSIR (Clontech), pIRES-EGFP (Clontech), pEAK-10 (Edge Biosystems) pTriEx-Hygro (Novagen) and pCINeo (Promega). Non-limiting examples for plasmid vectors suitable for *Pichia pastoris* comprise e.g. the plasmids pAO815, pPIC9K and pPIC3.5K (all Invitrogen).

Generally, vectors can contain one or more origins of replication (ori) and inheritance systems for cloning or expression, one or more markers for selection in the host, e.g., antibiotic resistance, and one or more expression cassettes. Examples of suitable origins of replication include, for example, the full length ColE1, its truncated versions such as those present on the pUC plasmids, the SV40 viral and the M13 phage origins of replication. Non-limiting examples of selectable markers include ampicillin, chloramphenicol, tetracycline, kanamycin, dhfr, gpt, neomycin, hygromycin, blasticidin or geneticin.

Further, said vector comprises a regulatory sequence that is operably linked to said nucleotide sequence or the nucleic acid molecule defined herein.

The coding sequence(s), e.g., said nucleotide sequence encoding the PA-rich polypeptide,

- 70 -

comprised in the vector can be linked to (a) transcriptional regulatory element(s) and/or to other amino acid encoding sequences using established methods. Such regulatory sequences are well known to those skilled in the art and include, without being limiting, regulatory sequences ensuring the initiation of transcription, internal ribosomal entry sites (IRES) and, optionally, regulatory elements ensuring termination of transcription and stabilization of the transcript. Non-limiting examples for such regulatory sequences ensuring the initiation of transcription comprise promoters, a translation initiation codon, enhancers, insulators and/or regulatory elements ensuring transcription termination. Further examples include Kozak sequences and intervening sequences flanked by donor and acceptor sites for RNA splicing, nucleic acid sequences encoding secretion signals or, depending on the expression system used, signal sequences capable of directing the expressed protein to a cellular compartment or to the culture medium.

Examples of suitable promoters include, without being limiting, the cytomegalovirus (CMV) promoter, SV40 promoter, RSV (Rous sarcome virus) promoter, the lacZ promoter, chicken β -actin promoter, CAG promoter (a combination of chicken β -actin promoter and cytomegalovirus immediate-early enhancer), human elongation factor 1 α promoter, AOX1 promoter, GAL1 promoter, CaM-kinase promoter, the lac, trp or tac promoter, the lacUV5 promoter, the T7 or T5 promoter, the *Autographa californica* multiple nuclear polyhedrosis virus (AcMNPV) polyhedral promoter or a globin intron in mammalian and other animal cells. One example of an enhancer is, e.g., the SV40 enhancer. Non-limiting additional examples for regulatory elements/sequences ensuring transcription termination include the SV40 poly-A site, the tk poly-A site or the AcMNPV polyhedral polyadenylation signals.

Furthermore, depending on the expression system, leader sequences capable of directing the polypeptide to a cellular compartment or secreting it into the medium may be added to the coding sequence of the nucleic acid molecule provided herein. The leader sequence(s) is (are) assembled in frame with translation, initiation and termination sequences, and preferably, a leader sequence is capable of directing secretion of translated protein, or a portion thereof, into the periplasmic space or into the extracellular medium. Suitable leader sequences are, for example, the signal sequences of BAP (bacterial alkaline phosphatase), CTB (cholera toxin subunit B), DsbA, ENX, OmpA, PhoA, stII, OmpT, PelB, Tat (Twin-arginine translocation) in *E. coli*, and the signal sequences of bovine growth hormone, human chymotrypsinogen, human factor VIII, human ig-kappa, human insulin, human interleukin-2, luciferase from *Metrida* or *Vargula*, human trypsinogen-2, inulinase from *Kluyveromyces marxianus*, mating factor alpha-1 from *Saccharomyces cerevisiae*, mellitin, human azurocidin and the like in eukaryotic cells.

- 71 -

The vectors may also contain an additional expressible nucleic acid sequence coding for one or more chaperones to facilitate correct protein folding.

Preferably, the vector of the present invention is an expression vector. An expression vector according to this invention is capable of directing the replication and the expression of the nucleic acid molecule of the invention, e.g., the nucleic acid molecule comprising the nucleotide sequence encoding the proline/alanine-rich polypeptide and the nucleotide sequence encoding the biologically active protein. In the appended examples, an expression vector comprising (i) a nucleotide sequence encoding a polypeptide consisting of proline, alanine and serine and (ii) a biologically active protein, such as IL-1Ra, was constructed; see Example 6. An exemplary expression vector comprising the nucleic acid molecule encoding a polypeptide consisting of proline and alanine is shown in Example 10.

Suitable bacterial expression hosts comprise, e.g., strains derived from *Escherichia coli* JM83, W3110, KS272, TG1, BL21 (such as BL21(DE3), BL21(DE3)PlysS, BL21(DE3)RIL, BL21(DE3)PRARE), Origami (K-12), Origami B or Rosetta. For vector modification, PCR amplification and ligation techniques, see methods described in Sambrook (2001) loc. cit..

Additionally, baculoviral systems can also be used as a vector in order to express the nucleic acid molecules of the invention in eukaryotic expression systems. In these aspects, the pFBDM vector can be used as an expression vector. The insertion into the MultiBac baculoviral DNA is mediated via the Tn7 transposition sequence upon transformation of DH10 MultiBac *E. coli* cells (Berger (2013) J. Vis. Exp. 77:50159, Fitzgerald (2006) Nat. Methods. 2006 3:1021-1032.). Virus amplification and expression can be performed in Sf21 (*Spodoptera frugiperda*) or High Five (*Trichoplusia ni*) cells.

The nucleic acid molecules and/or vectors of the invention as described herein above may be designed for introduction into cells by, e.g., non-chemical methods (electroporation, sonoporation, optical transfection, gene electrotransfer, hydrodynamic delivery or naturally occurring transformation upon contacting cells with the nucleic acid molecule of the invention), chemical-based methods (calcium phosphate, DMSO, PEG, liposomes, DEAE-dextrane, polyethylenimine, nucleofection etc.), particle-based methods (gene gun, magnetofection, impalefection), phage or phagemid vector-based methods and viral methods. For example, expression vectors derived from viruses such as retroviruses, vaccinia virus, adeno-associated virus, herpes viruses, Semliki Forest Virus or bovine papilloma virus, may be used for delivery of the nucleic acid molecules into a targeted cell population.

Preferably, the nucleic acid molecules and/or vectors of the invention are designed for

- 72 -

transformation of electrocompetent *E. coli* by electroporation or for stable transfection of CHO cells by calcium phosphate, polyethylenimine or lipofectaminetransfection (Pham (2006) Mol. Biotechnol. 34:225-237; Geisse(2012) Methods Mol. Biol. 899:203-219; Hacker (2013) Protein Expr. Purif. 92:67-76).

The present invention also relates to a host cell or a non-human host transformed with a vector or the nucleic acid molecule of this invention. It will be appreciated that the term "host cell or a non-human host transformed with the vector of the invention", in accordance with the present invention, relates to a host cell or a non-human host that comprises the vector or the nucleic acid molecule of invention. Host cells for the expression of polypeptides are well known in the art and comprise prokaryotic cells as well as eukaryotic cells. Thus, the host can be selected from the group consisting of a bacterium, a mammalian cell, an algal cell, a ciliate, yeast and a plant cell.

Typical bacteria include *Escherichia*, *Corynebacterium (glutamicum)*, *Pseudomonas (fluorescens)*, *Lactobacillus*, *Streptomyces*, *Salmonella* *Bacillus* (such as *Bacillus megaterium* or *Bacillus subtilis*), or *Corynebacterium* (like *Corynebacterium glutamicum*). The most preferred bacterium host herein is *E. coli*. An exemplary ciliate to be used herein is *Tetrahymena*, e.g. *Tetrahymena thermophila*.

Typical mammalian cells include, Hela, HEK293, HEK293T, H9, Per.C6 and Jurkat cells, mouse NIH3T3, NS0 and C127 cells, COS 1, COS 7 and CV1, quail QC1-3 cells, mouse L cells, mouse sarcoma cells, Bowes melanoma cells and Chinese hamster ovary (CHO) cells. Most preferred mammalian host cells in accordance with the present invention are CHO cells. An exemplary host to be used herein is *Cricetulus*, e.g. *Cricetulus griseus* (Chinese hamster). Also, human embryonic kidney (HEK) cells are preferred.

Other suitable eukaryotic host cells are e.g. yeasts such as *Pichia pastoris*, *Kluyveromyces lactis*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* or chicken cells, such as e.g. DT40 cells. Insect cells suitable for expression are e.g. *Drosophila* S2, *Drosophila* Kc, *Spodoptera* Sf9 and Sf21 or *Trichoplusia* Hi5 cells. Preferable algal cells are *Chlamydomonas reinhardtii* or *Synechococcus elongatus* cells and the like. An exemplary plant is *Physcomitrella*, for example *Physcomitrella patens*. An exemplary plant cell is a *Physcomitrella* plant cell, e.g. a *Physcomitrella patens* plant cell.

Also within the scope of the present invention are primary mammalian cells or cell lines. Primary cells are cells which are directly obtained from an organism. Suitable primary cells are, for example, mouse embryonic fibroblasts (MEF), mouse primary hepatocytes, cardiomyocytes

- 73 -

and neuronal cells as well as mouse muscle stem cells (satellite cells), human dermal and pulmonary fibroblasts, human epithelial cells (nasal, tracheal, renal, placental, intestinal, bronchial epithelial cells), human secretory cells (from salivary, sebaceous and sweat glands), human endocrine cells (thyroid cells), human adipose cells, human smooth muscle cells, human skeletal muscle cells, human leucocytes such as B-cells, T-cells, NK-cells or dendritic cells and stable, immortalized cell lines derived thereof (for example hTERT or oncogene immortalized cells). Appropriate culture media and conditions for the above described host cells are known in the art.

The host cells may e.g. be employed to produce large amounts of the nucleic acid molecule provided herein, the encoded polypeptide provided herein, and/or said drug conjugate provided herein. Hence, the host cells may be employed to produce large amounts of the nucleic acid molecule provided herein, the polypeptide encoded by the nucleic acid molecule comprising the nucleotide sequence encoding the PA-rich polypeptide and/or the polypeptide encoded by the nucleic acid molecule comprising the nucleotide sequence encoding the PA-rich polypeptide and the biologically active protein.

Accordingly, the present invention also relates to a method for preparing the nucleic acid molecule or the vector provided herein, the method comprising culturing the host or host cell of the invention under suitable conditions and optionally isolating the produced nucleic acid molecule and/or vector.

Furthermore, the present invention relates to a method for preparing a polypeptide encoded by the nucleic acid molecule or the nucleotide sequence provided herein, the method comprising culturing the host or host cell of the invention under suitable conditions and optionally isolating the produced polypeptide.

Furthermore, the present invention relates to a method for preparing a drug conjugate, the method comprising culturing the host cell of the invention under suitable conditions and optionally isolating the produced said drug conjugate. Preferably, the present invention relates to a method for preparing a drug conjugate, wherein said drug conjugate is encoded by the nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding a polypeptide consisting of proline, alanine and, optionally, serine and a biologically active protein.

Thus, the present invention relates in one aspect to a method for preparing a drug conjugate, wherein said drug conjugate comprises the polypeptide encoded by the nucleic acid molecule as defined herein and further comprises (i) a biologically active protein and/or (ii) a small

- 74 -

molecule and/or (iii) a carbohydrate, wherein the method further comprises culturing the host or host cell as provided herein and optionally isolating the produced polypeptide and/or drug conjugate. For example, if the drug conjugate is a fusion protein comprising the polypeptide encoded by the nucleic acid molecule as defined herein and further comprising a biologically active protein, the method can further comprise culturing the host or host cell as provided herein (i.e. a host or host cell comprising a nucleic acid encoding a polypeptide consisting of proline, alanine and, optionally, serine, as provided herein and the nucleic acid encoding a biologically active protein as defined herein, particularly a therapeutically active protein), and optionally isolating the produced fusion protein (drug conjugate). Of course, if the drug conjugate is a protein, the method can also comprise culturing the host or host cell as provided herein (i.e. a host or host cell comprising a nucleic acid encoding a polypeptide consisting of proline, alanine and, optionally, serine, as provided herein) and/or culturing the host or host cell comprising a nucleic acid encoding a biologically active protein as defined herein, particularly a therapeutically active protein, and optionally isolating the produced polypeptide consisting of proline, alanine and, optionally, serine and/or isolating the produced biologically active protein, and further optionally conjugating the polypeptide and the biologically active protein (e.g. by chemical coupling) to produce the drug conjugate.

For example, if the drug conjugate is a conjugate of a polypeptide encoded by the nucleic acid molecule as defined herein (i.e. a nucleic acid encoding a polypeptide consisting of proline, alanine and, optionally, serine, as provided herein) and of a small molecule and/or of a carbohydrate, the method can further comprise culturing the host or host cell as provided herein (i.e. a host or host cell comprising a nucleic acid encoding a polypeptide consisting of proline, alanine and, optionally, serine, as provided herein), and optionally isolating the produced polypeptide, and further optionally conjugating the polypeptide to the small molecule and/or carbohydrate (e.g. by chemical coupling).

“Culturing the host or host cell” includes in this context expression of the polypeptide as defined herein and/or of the biologically active protein in the host or host cell.

It is demonstrated in the appended examples that such a nucleic acid molecule comprising (i) a nucleotide sequence encoding a polypeptide consisting of proline, alanine and serine and (ii) a biologically active protein, such as IL-1Ra, can be bacterially expressed and, subsequently, be purified; see Example 8 and Figure 7. Furthermore, it is shown herein that a polypeptide consisting of proline and alanine encoded by the nucleic acid molecule provided herein can be expressed and purified; see e.g. Example 11 and Figure 8. By conjugation of the encoded polypeptide consisting of proline, alanine and, optionally, serine to a small molecule drug, a carbohydrate and/or a biologically active protein, the plasma half-life and/or solubility of the small molecule/small molecule drug and/or biologically active protein may be increased, unspecific toxicity may be decreased, and the prolonged exposure of the active drug to target

- 75 -

cells or structures in the body may result in enhanced pharmacodynamics.

The vector present in the host of the invention is either an expression vector, or the vector mediates the stable integration of the nucleic acid molecule of the present invention into the genome of the host cell in such a manner that expression of the protein is ensured. Means and methods for selecting a host cell in which the nucleic acid molecule of the present invention has been successfully introduced such that expression of the protein is ensured are well known in the art and have been described (Browne (2007) Trends Biotechnol. 25:425-432; Matasci (2008) Drug Discov. Today: Technol. 5:e37-e42; Wurm (2004) Nat. Biotechnol. 22:1393-1398).

Suitable conditions for culturing prokaryotic or eukaryotic host cells are well known to the person skilled in the art. For example, bacteria such as e.g. *E. coli* can be cultured under aeration in Luria Bertani (LB) medium, typically at a temperature from 4 to about 37°C. To increase the yield and the solubility of the expression product, the medium can be buffered or supplemented with suitable additives known to enhance or facilitate both. In those cases where an inducible promoter controls the nucleic acid molecule of the invention in the vector present in the host cell, expression of the polypeptide can be induced by addition of an appropriate inducing agent, such as, e.g., isopropyl- β -D-thiogalactopyranoside (IPTG) or anhydrotetracycline (aTc) as employed in the appended examples. Suitable expression protocols and strategies have been described in the art, e.g. in Sambrook (2001) loc. cit., (Gebauer (2012) Meth. Enzymol. 503:157-188) and can be adapted to the needs of the specific host cells and the requirements of the protein to be expressed, if required.

Depending on the cell type and its specific requirements, mammalian cell culture can, e.g., be carried out in RPMI, Williams' E or medium DMEM containing 10% (v/v) FCS, 2 mM L-glutamine and 100 U/ml penicillin/streptomycin. The cells can be kept, e.g., at 37°C, or at 41°C for DT40 chicken cells, in a 5% CO₂, water-saturated atmosphere. A suitable medium for insect cell culture is, e.g., TNM + 10% FCS, SF900 or HyClone SFX-Insect medium. Insect cells are usually grown at 27°C as adhesion or suspension cultures. Suitable expression protocols for eukaryotic or vertebrate cells are well known to the skilled person and can be retrieved, e.g., from Sambrook (2001) (loc. cit).

Preferably, the method for preparing the nucleic acid molecule, the vector, the polypeptide and/or the drug conjugate of the invention is carried out using either bacterial cells, such as, e.g., *E. coli* cells, or mammalian cells, such as, e.g., CHO cells. More preferably, the method is carried out using *E. coli* cells or CHO cells and most preferably, the method is carried out using *E. coli* cells.

Methods for the isolation of the encoded polypeptides produced comprise, without limitation, purification steps such as affinity chromatography (preferably using a fusion tag such as the *Strep*-tag II or the His₆-tag), gel filtration (size exclusion chromatography), anion exchange chromatography, cation exchange chromatography, hydrophobic interaction chromatography, high pressure liquid chromatography (HPLC), reversed phase HPLC, ammonium sulfate precipitation or immunoprecipitation. These methods are well known in the art and have been generally described, e.g., in Sambrook (2001) loc. cit. and are also described in the appended examples, see e.g. Examples 8 and 11. Such methods provide substantially pure polypeptides. Said pure polypeptides have a homogeneity of, preferably, at least about 90 to 95 % (on the protein level), more preferably, at least about 98 to 99 %. Most preferably, these pure polypeptides are suitable for pharmaceutical use/applications. Also, the application in food or cosmetic industry is envisaged herein. Depending upon the host cell / organism employed in the production procedure, the encoded polypeptides of the present invention may be glycosylated or may be non-glycosylated. Preferably, the polypeptide consisting of proline, alanine and, optionally, serine encoded by the nucleic acid molecule of the invention is not glycosylated. Most preferably, the polypeptide consisting of proline, alanine and, optionally, serine encoded by the nucleic acid molecule of the invention is not posttranslationally modified in its side chains such as, for example, by proline hydroxylation.

The encoded random coil polypeptide consists predominantly of alanine, proline and, optionally, serine residues, whereas serine, threonine or asparagine, which is required for O- or N-glycosylation, is preferably absent. Thus, the production of the polypeptide itself or of a biologically active protein comprising the encoded random coil polypeptide can result in a monodisperse product preferably devoid of post-translational modifications within the Pro/Ala/Ser or Pro/Ala sequence. This is an advantage for recombinant protein production in eukaryotic cells, like chinese hamster ovarian cells (CHO), HEK cells, or yeast, which are often chosen for the biosynthesis of complex proteins.

The invention also relates to a method for preparing a drug conjugate, wherein said drug conjugate comprises the polypeptide encoded by the inventive nucleic acid molecule provided herein and further comprises (i) a biologically active protein and/or (ii) a small molecule and/or (iii) a carbohydrate. Such carbohydrate conjugates may be particularly useful as vaccines.

As described above, a drug conjugate comprising the PA-rich polypeptide and the biologically active protein can be prepared by expressing the nucleic acid molecule comprising the nucleotide sequence encoding the PA-rich polypeptide and the nucleic acid sequence encoding the biologically active protein. The expressed drug conjugate can be isolated. Alternatively, the

- 77 -

drug conjugate can be prepared by culturing/raising the host comprising the nucleotide sequence or the nucleic acid molecule encoding said polypeptide consisting of proline, alanine and, optionally serine. Thus, the nucleic acid molecule is expressed in the host. Optionally, said produced polypeptide is isolated. The produced polypeptide consisting of proline, alanine and, optionally, serine can be conjugated to the biologically active protein, e.g., via a peptide bond or a non-peptide bond.

It is demonstrated in the illustrative examples that the PA-rich polypeptide encoded by the nucleic acid molecule provided herein can be expressed in bacteria and can be purified therefrom; see e.g. Example 11 and Figure 8. In particular, it was surprisingly shown that the start methionine (N-terminal methionine) of the prepared polypeptide is cleaved off and is thus missing in the produced polypeptide; see e.g. Example 12 and Figure 8D. The missing start methionine in the prepared polypeptide allows the seamless conjugation of the primary amino group of the following amino acid (after the start methionine) at the N-terminus to a biologically active protein, a small molecule and/or a carbohydrate. Therefore, the nucleic acid molecules of the invention and the prepared polypeptides thereof are particularly advantageous for the conjugation to, e.g., the biologically active proteins.

In particular, the biologically active protein can be site-specifically conjugated, e.g., in the presence of an activating agent such as N-(3-dimethylaminopropyl)-N'-ethylcarbodiimide (EDC) or as an N-hydroxysuccinimide (NHS) ester (Hermanson (1996) *Bioconjugate Techniques*, 1st edition, Academic Press, San Diego, CA) to the N-terminus of the produced random coil polypeptide. Alternatively, the biologically active protein can be site-specifically conjugated to the C-terminus of the produced random coil polypeptide consisting of proline, alanine and, optionally, serine, e.g., in the presence of an activating agent such as EDC or after activation as an NHS ester.

Furthermore, the produced polypeptide consisting of proline, alanine and, optionally, serine can be conjugated to the small molecule and/or to the carbohydrate via a non-peptide bond. Non-peptide bonds that are useful for cross-linking proteins are known in the art and may include disulfide bonds, e.g., between two Cys side chains and/or thiol groups, thioether bonds and amide bonds between carboxyl groups and amino groups. Non-peptide covalent bonds may also be provided by chemical cross-linkers, such as disuccinimidyl suberate (DSS), N- β -maleimidopropyl-oxysuccinimide ester (BMPS) or sulfosuccinimidyl 4-[p-maleimidophenyl]butyrate (Sulfo-SMPB), metal-chelating/complexing groups as well as non-covalent protein-protein or protein-peptide interactions.

Furthermore, a small molecule drug can be site-specifically conjugated to the polypeptide

- 78 -

forming the random coil. Optionally, the N-terminus of the polypeptide can be modified with a suitable protective group, for example an acetyl group or a pyroglutamyl group, and after activation of the C-terminal carboxylate group, e.g. using the common reagents EDC and NHS (Hermanson (1996) loc. cit.), site-specific coupling of the drug to the C-terminus of the random coil polypeptide can be achieved. In this manner uniform drug conjugates can be easily obtained.

As an alternative to a single site-specific conjugation the random coil polypeptide consisting of proline, alanine and, optionally, serine may be equipped with additional side chains, at the N- or the C-terminus or internally, suitable for chemical modification, such as lysine residues with their ϵ -amino groups, cysteine residues with their thiol groups, or even non-natural amino acids, allowing the conjugation of one, two or multiple small molecules using, for example, NHS or maleimide active groups.

Apart from stable conjugation, a prodrug may be linked transiently to the random coil polypeptide. The linkage can be designed to be cleaved *in vivo*, in a predictable fashion, either via an enzymatic mechanism or by slow hydrolysis initiated at physiological pH similarly as, for example, the poorly soluble antitumor agent camptothecin was conjugated to a PEG polymer, thus achieving increased biodistribution, decreased toxicity, enhanced efficacy and tumor accumulation (Conover (1998) Cancer Chemother. Pharmacol. 42:407-414). Examples for further prodrugs are chemotherapeutic agents like docetaxel (Liu (2008) J. Pharm. Sci. 97:3274-3290), doxorubicin (Veronese (2005) Bioconjugate Chem. 16: 775-784) or paclitaxel (Greenwald (2001) J. Control Release 74:159-171).

It is also envisaged herein that the small molecule may be coupled to a fusion protein, e.g., the polypeptide forming the random coil consisting of proline, alanine and, optionally, serine genetically fused to a targeting domain, e.g. an antibody fragment, thus resulting in specific delivery of the small molecule drug. The immunotoxin generated in the latter case by conjugation with a cytotoxic small molecule is particularly useful if the targeting domain is directed against a cell-surface receptor which undergoes internalization, for example.

As used herein, the term "drug" refers to a small molecule, a biologically active protein, a peptide or a carbohydrate. As used herein, the term "small molecule" can refer to an (organic) low molecular weight (<900 Daltons) compound. Small molecules can help to regulate a biological process and have usually a size in the order of nanometer. It is envisaged herein that the small molecule is used in a method of therapy, diagnosis or is used in the food or cosmetic industry. For example, the drug conjugate with the produced polypeptide that is encoded by the nucleotide sequence or the nucleic acid molecule provided herein can comprise (a) small

- 79 -

molecule(s) that is/are selected from the group consisting of angiogenesis inhibitors, anti-allergic drugs, anti-emetic drugs, anti-depressant drugs, anti-hypertensive drugs, anti-inflammatory drugs, anti-infective drugs, anti-psychotic drugs, anti-proliferative (cytotoxic and cytostatic) drugs, calcium antagonists and other circulatory organ drugs, cholinergic agonists, drugs acting on the central nervous system, drugs acting on the respiratory system, steroids, antisense nucleic acids, small interference RNAs (siRNAs), micro RNA (miR) inhibitors, microRNA mimetics, DNA aptamers and RNA aptamers.

Exemplary angiogenesis inhibitors include but are not limited to MetAP2 inhibitors (like fumagillin, fumagillin derivatives, 2-{3-[3,5-bis[4-nitrobenzylidene]-4-oxopiperidin-1-yl]-3-oxopropylsulfanyl} ethanesulfonic acid), VGFR inhibitors (like axitinib, brivanib, cabozantinib, tivozanib and motesanib), placenta growth factor (PIGF) inhibitors, platelet-derived growth factor receptor inhibitors (like AC 710, sorafenib, sunitinib, and vatalanib) and the like.

Exemplary anti-allergic drugs include but are not limited to antihistaminics (like diphenhydramine (benadryl), dimenhydrinate (dramamine, driminate), hydroxyzine hydrochloride (restall, vistacot), promethazine (phenergan)) and the like.

Exemplary anti-depressant drugs include but are not limited to granisetron, palonosetron, and the like.

Exemplary anti-depressant drugs include but are not limited to cis-flupenthixol, imipramine hydrochloride, mianserin and the like.

Exemplary anti-hypertensive drugs include but are not limited to alprostadil, diazoxide, nicardipine and the like.

Exemplary anti-inflammatory drugs include but are not limited to, cortisone, hyaluronic acid, ketorolac and the like.

Exemplary anti-infective drugs include but are not limited to aminoglycosides, amadovir, amoxicillin, ampicillin, benzylpenicillin, carbapenems, cephalosporin, ceftiofur, chloramphenicol, cefepime, ceftazidime, ceftobiprole, clindamycin, draxxin, dalbavancin, daptomycin, dihydrostreptomycin, erythromycin, florfenicol, fluoroquinolones, flunixin meglumine, linezolid, marbofloxacin, micafungin, nitrofurazone, oritavancin, oxytetracycline, penicillin, piperacillin, procain, rupintrivir, spiramycin, streptogramins, sulfadimethoxin, sulfamethazine, tedizolid, telavancin, ticarcillin, tilmicosin, tigecycline, tildipirosin, tylosin,

- 80 -

vancomycin, and the like.

Exemplary anti-psychotic drugs include but are not limited to amisulprid, ariprazole, benperidol, bromperidol, clorpromazin, chlorprothixen, clopenthixol, clozapine, flupentixol, fluphenazin, fluspirilen, haloperidol, levomepromazin, melperon, olanzapine, perazin, perphenazin, pimozid, pipamperon, promazin, promethazine, prothipendyl, quetiapine, risperido, sulpirid, thioridazin, trifluoperazin, triflupromazin, zuclopenthixol, and the like.

Exemplary anti-tumor drugs include but are not limited to anthracyclins (like doxorubicin, epirubicin, idarubicin, and daunorubicin), alkylating agents (like calicheamicins, dactinomycines, mitromycines, and pyrrolbenzodiazepines), AKT inhibitors (like AT7867), amatoxins (like ax-amanitins, P-amanitins, y-amanitins, c-amanitins, amanullin, amanullic acid, amaninamide, amanin, and proamanullin, SN-38, and camptothecin), ATM inhibitors, auristatins (like auristatin EB (AEB), auristatins EFP (AEFP), monomethyl auristatin E (MMAE), monomethyl auristatin F (MMAF), auristatin F, and dolastatin), cryptophycins, cyclin-dependent kinases inhibitors (like BMS-387032, PD0332991, GSK429286, AZD7762; AZD 1152, MLN8054 and MLN8237; BI 2536, B16727, GSK461364, ON-01910, SB 743921, SB 715992, MK-0731, AZD8477, AZ3146, and ARRY-520), duocarmycins, DNA-PK inhibitors, epothilones (like epothilone A, B, C, D, E or F, and derivatives), GSK-3 inhibitors, HDAC inhibitors (like belinostat, CUDC- 101, droxinostat, ITF2357, JNJ-26481585, LAQ824, and panobinostat MC1568, mocetinostat, entinostat, PCI-24781, pyroxamide, trichostatin A, and vorinostat), hsp70 inhibitors, hsp90 inhibitors (like 17AAG derivatives, B11B021, B11B028, SNX-5422, NVP-AUY-922, KW-2478, and geldanamycin), MAPK signaling pathway inhibitors (like MEK, Rac, JNK, B-Raf), maytansinoids, maytansinoid analogs (like maytansinol, maytansinol analogs, maytansine, DM-1, and DM-4), p38 MAPK inhibitors (like GDC-0973, GSK1 120212, MSC1936369B, AS703026, R05126766 and R04987655, PD0325901, AZD6244, AZD 8330, GDC-0973, CDC-0879, PLX-4032, SB590885, BIRB 796, LY2228820, SB 202190, AEE788, BIBW2992, afatinib, lapatinib, erlotinib, and gefitinib), PARP inhibitors (like iniparib, olaparib, veliparib, AG014699, CEP 9722, MK 4827, KU-0059436, LT-673, 3 aminobenzamide, A-966492, and AZD2461), PDK-1 inhibitors, platinum compounds (like cisplatin, carboplatin, oxaliplatin, iproplatin, ormaplatin or tetraplatin), taxans (like paclitaxel, and ordocetaxel), tubulysins (like tubulysin A, tubulysin B and tubulysin derivatives), vinca alkaloids (like vinblastine, vindesine and navelbine), Wnt/Hedgehog signaling pathway inhibitors like (vismodegib, GDC-0449, cyclopamine, and XAV-939), and the like.

Exemplary drugs acting on the central nervous system include but are not limited to buprenorphin, cryostatin, naltroxrexone, naloxone, and the like.

Exemplary vitamins include but are not limited to vitamin B-12 (cyanocobalamin), vitamin A, and the like.

Exemplary steroids include but are not limited to androgenic steroids (like fluoxymesterone, methyltestosterone, testosterone, trenbolone), estrogens (like beta-estradiol, diethylstilbestrol, estrone, estriol, equilin, estropipate equilin, mestranol), progestational compounds (like 19-norprogesterone, alfaprostol, chlormadinone, demegestone, dydrogesterone, dimethisterone, ethisterone, ethynodiol diacetate, norethindrone, norethindrone acetate, medroxyprogesterone, melengestrolprogesterone, norgestrel, promegestone, zeranol), and the like.

Exemplary antisense nucleic acids include but are not limited to antisense nucleic acids targeting the androgen receptor (like ISIS-AR, AZD5312), angiopoietin-like 3 protein (like ISIS-ANGPTL3), apolipoprotein B100 (like mipomersen), apolipoprotein CIII (like ISIS-APOCIII, volanesorsen), the connective tissue growth factor (CTGF) (like EXC 001, PF-06473871), clusterin (like custirsen, OGX-011), C- reactive protein (like ISIS-CRP), diacylglycerol acyltransferase (like ISIS-DGAT2), factor VII (like ISIS-FVII), fibroblast growth factor receptor 4 (like ISIS-FGFR4), hepcidin (like XEN701), Hsp27 (like apatorsen, OGX-427), the HTT gene (like ISIS-HTT), ICAM-1 (like alicaforsen), prekallikren (like ISIS-PKK), SMN2 (like ISIS-SMN), STAT3 (like ISIS-STAT3-2.5, AZD9150), the transthyretin gene (like ISIS-TTR), and the like.

Exemplary small interference RNAs (siRNAs) include but are not limited to siRNAs targeting the alpha-1-antitrypsin mutant Z-AAT (like ALN-AAT), aminolevulinate synthase 1 (ALAS-1) (like ALN-AS1, ALN-AS2), antithrombin III (like ALN-AT3), the complement component C5 (like ALN-CC5), the complement component C6 (like ALN-CC6), the connective tissue growth factor (like RXI-109), exon 8 of the dystrophin gene (like SRP-4008), exon 44 of the dystrophin gene (like SRP-4044), exon 45 of the dystrophin gene (like SRP-4045), exon 50 of the dystrophin gene (like SRP-4050), the ebola virus (like AVI-7537), exon 51 of the dystrophin gene (like eteplirsen, AVI-4658), exon 52 of the dystrophin gene (like SRP-4052), exon 53 of the dystrophin gene (like SRP-4053), the influenza virus (like AVI-7100), the kinesin spindle protein (KSP), lung diseases (like Atul11), the marburg virus (like AVI-7288), multi-targeted small interfering RNA (siRNA) cocktails (like STP503, STP523, STP601, STP702, STP705, STP801, STP805, STP900, STP902, STP911, STP916, siPOOLS), the nucleocapsid N of the virus genome (like ALN-RSV01), PCSK9 (like ALN-PCS01, ALN-PCSSc), the pro-apoptotic protein caspase 2 (like QPI-1007), the pro-apoptotic protein p53 (like QPI-1002), RTP801 (like PF-655), SERPINC1 (like ALN-AT4), the transmembrane protease serine 6 (Tmprss6) (like ALN-TMP), transthyretin (like ALN-TTRsc, ALN-TTR02), PCSK10

- 82 -

(like ALN-PCS02), PKN3 (like Atu027), the vascular endothelial growth factor (VEGF) (like ALN-VSP), and the like.

Exemplary microRNA inhibitors include but are not limited to inhibitors of miR-10b, miR-15, miR-21, miR-29, miR-33, miR-92, miR-145, miR195, miR-208, miR-221, miR-451, miR-499 and the like.

Exemplary microRNA mimetics include but are not limited to an analog of miR-34 (like MRX34), miR-Rx06, miR-Rx07, miR-Rx16, an analog of let7 (like miR-Rxlet-7), and the like. Exemplary DNA aptamers include but are not limited to nucleolin inhibitor (like AS1411), pGDF inhibitor (like E10030), thrombin inhibitor (like NU172), vWF inhibitor (like ARC1779), and the like.

Exemplary RNA aptamers include but are not limited to C5a inhibitor (like NOX-D21 or ARC1905), calcitonin gene-related peptide inhibitor (like NOX-L41), C-C chemokine ligand 2 inhibitor (like NOX-E36), CXCL12 inhibitor (like NOX-A12), glucagon inhibitor (like NOX-G16), hepcidin antagonist (like NOX-H94), pathogen recognition receptor agonist (like a RIG-I agonist), sphingosine-1-phosphate inhibitor (like NOX-S93), VEGF antagonist (like NX1838), and the like.

Exemplary carbohydrates that are potentially useful for the preparation of vaccines include but are not limited carbohydrate epitopes specifically bound by lectins, *E. coli* O 121 O-antigen, *E. coli* O 121 O-antigen derivatives, Man9 from HIV-I gp120, shigella flexneri 2a O-polysaccharides, staphylococcus aureus polysaccharide capsular polysaccharide 5, staphylococcus aureus polysaccharide capsular polysaccharide 8, tumor-associated carbohydrate antigens (TACA) (like Tn antigens (e.g. α -GalNAc-O-Ser/Thr), sialyl Tn antigens (e.g. NeuAc α (2,6)-GalNAc α -O-Ser/Thr), Thomsen-Friedenreich antigen (Gal β 1-3GalNAc α 1), LewisY (e.g. Fuca(1,2)-Gal β (1,4)-[Fuca(1,3)]-GalNAc), sialyl LewisX and sialyl LewisA, LewisX (stage-specific embryonic antigen-1/SSEA-1), Globo H antigen (e.g. Fuca(1,2)-Gal β (1,3)-GalNAc β (1,3)-Gal α (1,4)-Gal β (1,4)-Glc), T antigen (e.g. Gal β (1,3)-GalNAc α -O-Ser/Thr), glycosphingolipid stage-specific embryonic antigen-3 (SSEA-3), sialic acid containing glycosphingolipids, ganglioside GD2, GD3, ganglioside GM2, ganglioside fucosyl GM and ganglioside Neu5GcGM3), and the like.

The drug conjugate comprising the polypeptide encoded by the inventive nucleic acid molecule provided herein comprising a biologically active protein and/or a small molecule and/or a carbohydrate may be used for the treatment of inflammatory diseases, infectious diseases, respiratory diseases, endocrine disorders, diseases of the central nervous system,

- 83 -

musculoskeletal diseases, cardiovascular diseases, oncological diseases, urogenital diseases and metabolic diseases.

Exemplary inflammatory diseases include but are not limited to ankylosing spondylitis, arthritis, atherosclerosis, atypical hemolytic uremic syndrome (aHUS), fibromyalgia, Guillain Barré syndrome (GBS), irritable bowel syndrome (IBS), Crohn's disease, colitis, dermatitis, diverticulitis, osteoarthritis, psoriatic arthritis, Lambert-Eaton myasthenic syndrome, systemic lupus erythematosus (SLE), nephritis, Parkinson's disease, multiple sclerosis, paroxysmal nocturnal hemoglobinuria (PNH), rheumatoid arthritis (RA), Sjögren's syndrome, ulcerative colitis, and the like.

Exemplary infectious diseases include but are not limited to African trypanosomiasis, borreliosis, cholera, cryptosporidiosis, dengue fever, hepatitis A, hepatitis B, hepatitis C, HIV/AIDS, influenza, Japanese encephalitis, leishmaniasis, malaria, measles, meningitis, onchocerciasis, pneumonia, rotavirus infection, schistosomiasis, sepsis, shigellosis, streptococcal tonsillitis, tuberculosis, typhoid, yellow fever, and the like.

Exemplary respiratory diseases include but are not limited to asthma, chronic obstructive pulmonary disease (COPD), cystic fibrosis, and the like.

Exemplary endocrine disorders include but are not limited to acromegaly, type I diabetes, type II diabetes, gestational diabetes, Graves' disease, growth hormone deficiency, hyperglycemia, hyperparathyroidism, hyperthyroidism, hypoglycemia, infertility, obesity, parathyroid diseases, Morquio A syndrome, mucopolysaccharidosis, and the like.

Exemplary diseases of the central nervous system include but are not limited to Alzheimer's disease, cataplexy, Huntington's disease, Parkinson's disease, and the like.

Exemplary musculoskeletal diseases include but are not limited to osteoporosis, muscular dystrophy, and the like.

Exemplary cardiovascular diseases include but are not limited to acute heart failure, cerebrovascular disease (stroke), ischemic heart disease, and the like.

Exemplary oncological diseases include but are not limited to adrenal cancer, bladder cancer, breast cancer, colon and rectal cancer, endometrial cancer, kidney cancer, acute lymphoblastic leukemia (ALL) and other types of leukemia, lung cancer, melanoma, Non-Hodgkin lymphoma, pancreatic cancer, prostate cancer, thyroid cancer, and the like.

Exemplary urogenital diseases include but are not limited to benign prostatic hyperplasia (BPH), hematuria, neurogenic bladder, Peyronie's disease, and the like.

Exemplary metabolic diseases include but are not limited to Gaucher disease, Fabry disease, Growth hormone deficiency, Hurler syndrome, Hunter syndrome, hyperoxaluria, neuronal ceroid lipofuscinosis, Maroteaux-Lamy syndrome, Morquio syndrome, Noonan syndrome, SHOX gene haploinsufficiency, Turner syndrome, Prader-Willi syndrome, phenylketonuria, Sanfilippo syndrome, , and the like.

As described above, the nucleic acid molecule provided herein can also be employed alone or as part of a vector for gene therapy purposes. Gene therapy, which is based on introducing therapeutic genes into cells by *ex vivo* or *in vivo* techniques, is one of the most important applications of gene transfer. Suitable vectors, methods or gene delivery systems for *in vivo* gene therapy are described in the literature and are known to the person skilled in the art; see, e.g., Giordano (1996) Nat. Med. 2:534-539; Schaper (1996) Circ. Res. 79:911-919; Anderson (1992) Science 256:808-813; Verma (1997) Nature 389:239-249; Isner (1996) Lancet 348:370-374; Muhlhauser (1995) Circ. Res. 77:1077-1086; Onodera (1998) Blood 91:30-36; Verma (1998) Gene Ther. 5:692-699; Nabel (1997) Ann. N.Y. Acad. Sci. 811:289-292; Verzeletti (1998) Hum. Gene Ther. 9:2243-2251; Wang (1996) Nat. Med. 2:714-716; WO 94/29469; WO 97/00957, US 5,580,859; US 5,589,466; or Schaper (1996) Curr. Opin. Biotechnol. 7:635-640. The nucleic acid molecules and vectors provided herein may be designed for direct introduction or for introduction via liposomes or viral vectors (e.g., adenoviral, retroviral) into the cell. For example, the vector can be an adeno-associated-virus (AAV) vector, in particular, an AAV8 vector. AAV vectors are attractive for gene therapy. The AAV system has several advantages including long-term gene expression, the inability to autonomously replicate without a helper virus, transduction of dividing and nondividing cells, and the lack of pathogenicity from wild-type infections. Preferably, said cell in which the nucleic acid molecule or vector is introduced is a germ line cell, embryonic cell or egg cell or derived therefrom, most preferably said cell is a stem cell. An example for an embryonic stem cell can be, inter alia, a stem cell as described in Nagy (1993) Proc. Natl. Acad. Sci. USA 90:8424-8428.

The terms "polypeptide", "peptide" and "protein" are used herein interchangeably and refer to a polymer of two or more amino acids linked via amide or peptide bonds that are formed between an amino group of one amino acid and a carboxyl group of another amino acid. Preferably, a peptide bond is formed between the α -amino group of one amino acid and the α -carboxyl group of another amino acid. The amino acids comprised in the peptide or protein, which are also referred to as amino acid residues, may be selected from the 20 standard

- 85 -

proteinogenic α -amino acids (i.e., Ala, Arg, Asn, Asp, Cys, Glu, Gln, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, and Val) but also from non-proteinogenic and/or non-standard α -amino acids (such as, e.g., ornithine, citrulline, homolysine, pyrrolysine, or 4-hydroxyproline) as well as β -amino acids (e.g., β -alanine), γ -amino acids and δ -amino acids. Preferably, the amino acid residues comprised in the peptide or protein are selected from α -amino acids, more preferably from the 20 standard proteinogenic α -amino acids (which can be present as the L-isomer or the D-isomer, and are preferably all, except for Gly, are present as the L-isomer).

The encoded polypeptide or protein may be unmodified or may be modified, e.g., at its N-terminus, at its C-terminus and/or at the side chain of any of its amino acid residues (particularly at the side chain functional group of one or more Lys, His, Ser, Thr, Tyr, Cys, Asp, Glu, and/or Arg residues). Such modifications may include, e.g., the attachment of any of the protecting groups described for the corresponding functional groups in: Wuts(2006) Greene's protective groups in organic synthesis, John Wiley & Sons, 4th edition, Hoboken, N.J.. Such modifications may also include glycosylation and/or acylation with one or more fatty acids (e.g., one or more C₈₋₃₀ alkanolic or alkenolic acids; forming a fatty acid acylated peptide or protein). The encoded polypeptide is preferably not hydroxylated, in particular does not comprise hydroxyproline.

The amino acid residues comprised in the encoded peptide or protein may, e.g., be present as a linear molecular chain (forming a linear peptide or protein) or may form one or more rings (corresponding to a cyclic peptide or protein), e.g., circularized via a peptide or isopeptide bond or a disulfide bridge. The peptide or protein may also form oligomers consisting of two or more identical or different molecules. As used herein, the term "domain" relates to any region/part of an amino acid sequence that is capable of autonomously adopting a specific structure and/or function. In the context of the present invention, accordingly, a "domain" may represent a functional domain or a structural domain, which may for example form part of a larger polypeptide.

As used herein, the terms "comprising" and "including" or grammatical variants thereof are to be taken as specifying the stated features, integers, steps or components but do not preclude the addition of one or more additional features, integers, steps, components or groups thereof. These terms encompass the terms "consisting of" and "consisting essentially of."

Thus, the terms "comprising"/"including"/"having" mean that any further component (or likewise features, integers, steps and the like) can/may be present. Thus, whenever the terms

- 86 -

“comprising”/“including”/“having” are used herein, they can be replaced by “consisting essentially of” or, preferably, by “consisting of”.

The term “consisting of” means that no further component (or likewise features, integers, steps and the like) is present.

The term “consisting essentially of” or grammatical variants thereof when used herein are to be taken as specifying the stated features, integers, steps or components but do not preclude the addition of one or more additional features, integers, steps, components or groups thereof but only if the additional features, integers, steps, components or groups thereof do not materially alter the basic and novel characteristics of the claimed composition, device or method.

Thus, the term “consisting essentially of” means that specific further components (or likewise features, integers, steps and the like) can be present, namely those not materially affecting the essential characteristics of the composition, device or method. In other words, the term “consisting essentially of” (which can be interchangeably used herein with the term “comprising substantially”), allows the presence of other components in the composition, device or method in addition to the mandatory components (or likewise features, integers, steps and the like), provided that the essential characteristics of the device or method are not materially affected by the presence of other components.

The term “method” refers to manners, means, techniques and procedures for accomplishing a given task including, but not limited to, those manners, means, techniques and procedures either known to, or readily developed from known manners, means, techniques and procedures, by practitioners of the chemical, biological and biophysical arts.

As used herein and if not indicated otherwise, the term “about” preferably refers to $\pm 10\%$ of the indicated numerical value, more preferably to $\pm 5\%$ of the indicated numerical value, and in particular to the exact numerical value indicated.

The present invention is further described by reference to the following non-limiting figures and examples. Unless otherwise indicated, established methods of recombinant gene technology were used as described, for example, in Sambrook (2001) *loc. cit.* which is incorporated herein by reference in its entirety.

The present invention is further described by reference to the following non-limiting figures and examples. The figures show:

Figure 1: Assembly of low repetitive nucleic acids encoding proline/alanine-rich sequences using the plasmid pXL2.

(A) Plasmid map of pXL2 (SEQ ID NO: 48). A *SapI* recognition site (5'-GCTCTTC-3') and an *EcoRI* recognition site (5'-CTCTTC-3') are inserted into the selectable marker gene *lacZ*, which is under transcriptional control of the *lac* promoter/operator (*lac^{P/O}*). The *SapI* restriction site allows the consecutive insertion of one or more units (building blocks) of low repetitive nucleic acid molecules encoding proline/alanine-rich sequences. Cutting with *EcoRI*, a restriction enzyme whose recognition sequence overlaps with / is a subset of the *SapI* recognition sequence, allows the excision and/or isolation of the assembled low repetitive nucleic acid molecule (gene cassette) encoding a proline/alanine-rich sequence. The plasmid backbone is identical with that of the generic cloning vector pUC19 (Yanisch-Perron (1985) Gene 33:103-119), except for deletion of additional *SapI* and *EcoRI* restriction sites via silent mutation. (B) Plasmid map of pXL1 (SEQ ID NO: 55). Two *SapI* recognition sites (5'-GCTCTTC-3') in opposite orientation are inserted into the selectable marker gene *lacZ*, which is under transcriptional control of the *lac* promoter/operator (*lac^{P/O}*). The *SapI* restriction site allows the insertion/cloning as well as propagation/amplification and the subsequent excision and/or isolation of a low repetitive nucleic acid molecule (gene cassette) encoding a proline/alanine-rich sequence. The plasmid backbone is identical with that of pUC19, except for deletion of an additional *SapI* restriction site via silent mutation. Note that the inserted gene cassette has reverse orientation compared with pXL2. (C) Nucleotide sequence stretch of and around the cloning site of pXL2 useful for the cloning and assembly of low repetitive nucleic acid molecules. The nucleotide sequence contains a *SapI* restriction site and an *EcoRI* restriction site in reverse orientation. As the *EcoRI* recognition sequence is also part of the *SapI* recognition sequence, *EcoRI* cuts at both recognition sites. Regardless of the restriction enzyme used, *SapI* or *EcoRI*; digest leads to protruding GCC/CGG ends (encoding Ala), which are compatible with the GCC/CGG overhangs of low repetitive nucleic acid molecules encoding proline/alanine-rich sequences and thus allow efficient ligation. Recognition sequences are underlined. (D) Nucleotide sequence and encoded amino acid sequence of the cloning site of pXL2 after insertion of one low repetitive nucleic acid unit/building block, PAS#1b(200) (SEQ ID NO: 19). The *SapI* and *EcoRI* restriction sites flanking the cloned low repetitive nucleic acid unit/building block are labelled (recognition sequences are underlined). (E) Assembly of low repetitive nucleic acid sequence units to obtain longer cloned nucleotide sequences (nucleic acid molecules) encoding proline/alanine-rich amino acid repeat sequences. In the first step, pXL2 is digested with *SapI*, dephosphorylated and ligated with the first sequence unit, PAS#1b(200). In the second step, the resulting plasmid is opened/linearized upstream of the cloned sequence unit by singular *SapI* restriction digest, followed by dephosphorylation and

- 88 -

ligation with the second nucleotide sequence unit, PAS#1c(200). The resulting plasmid pXL2-PAS#1c/1b(400) contains an inserted gene/DNA cassette comprising in total 1200 base pairs in length. Overall, the resulting cloned low repetitive nucleotide sequence encoding a proline/alanine-rich amino acid repeat sequence contains only nucleotide repeats with a maximum length of 14 nucleotides (SEQ ID NO: 52). The entire assembled gene cassette / nucleic acid molecule can be easily excised via digestion with *EcoRI* and subsequently used for subcloning on expression vectors, for example in the same reading frame with a nucleotide sequence encoding a biologically active protein, and the like. Notably, by repeating the second step, gene cassettes with successively increasing length can be assembled and cloned on pXL2 in a systematic fashion. If different suitable nucleotide sequence units are used, the resulting long low repetitive nucleic acid molecule encoding a proline/alanine-rich amino acid repeat sequence contains only few or short nucleotide repeats.

Figure 2: Repeat analysis of proline/alanine-rich sequences.

(A) Dot plots of the prior art proline/alanine-rich sequence PA#3a(200) (SEQ ID NO: 15) are compared to dot plots of a low repetitive nucleotide sequence according to this invention, PA#3b(200) (SEQ ID NO: 36), both generated using the dot plot tool "dottup" included in the Geneious V8.1 software package (Biomatters, Auckland, New Zealand), applying a repeat window of 14 or 15. By comparing the analyzed nucleotide sequence to itself on the x- and y-axes, respectively, of a two-dimensional graph the "dottup" tool identifies all regions where an identical sequence of specified length (repeat window) occurs and draws a diagonal line, thus indicating the positions of a repeat sequence on the x- and y-axis (if represented by a diagonal line different from the central diagonal line, the latter indicating self-identity). Successive repeats extend a diagonal line. The PA#3a(200) nucleotide sequence analyzed here reveals a highly repetitive nature as illustrated by the many and/or long diagonal lines. In this case, the nucleotide sequence shows numerous repeats of 60 base pairs each. By applying a repeat window of 14, even a shorter 14 bp repeat within the 60 bp repeat appears. In contrast, the PA#3b(200) nucleotide sequence shows only a few scattered 14 bp repeats within the entire analyzed nucleotide sequence of 600 bp whereas no repeat is detectable when applying a slightly larger repeat window of 15 for analysis. (B) Dot plots of the nucleic acid PAS#1a(600) (SEQ ID NO: 12) disclosed in WO2008155134 encoding a proline/alanine-rich sequence are compared to dot plots of the assembled low repetitive nucleotide sequences encoding proline/alanine-rich amino acid repeat sequences according to this invention, PAS#1d/1f/1c/1b(800) (SEQ ID NO: 39) and PAS#1f/1c/1b(600) (SEQ ID NO: 38), using repeat windows of 14 and 15 base pairs. Whereas the nucleotide sequence PAS#1a(600) reveals a composition of 60 base pair repeats, the PAS#1d/1f/1c/1b(800) and the PAS#1f/1c/1b(600) nucleotide sequences show no or, in the case of the 14 nucleotide repeat

- 89 -

window, only one single 14 nucleotide repeat (diagonal line) within the entire analyzed nucleotide sequences of 2400 or 1800 bp, respectively. (C) Dot plot analysis of a synthetic DNA sequence encoding an $[(AP)_5]_n$ multimer (SEQ ID NO: 16) disclosed in US2006/0252120 and of a natural DNA sequence encoding a repetitive proline/alanine-rich region of the very large tegument protein of *Macacine herpesvirus 1* (GenBank AAP41454.1) (SEQ ID NO: 18) in comparison with the low repetitive nucleotide sequence according to this invention, PA#3a(200) (SEQ ID NO: 15), applying repeat windows of 14 and 15 nucleotides. Dot plots of the DNA sequences encoding the $[(AP)_5]_n$ multimer and the proline/alanine-rich region of the very large tegument protein of *Macacine herpesvirus 1* illustrate the highly repetitive nature of these nucleotide sequences. In contrast, the dot plot of the PA#1b(200) nucleotide sequence shows only a few scattered 14 nucleotide repeats (14 nucleotide repeat window) or no repeat at all (15 nucleotide repeat window) within the entire analyzed nucleotide sequence of 600 bp.

Figure 3: Electropherogram from DNA sequencing of a low repetitive nucleic acid molecule encoding proline/alanine-rich amino acid sequences.

Electropherogram from DNA sequencing of pXL2-PAS#1f/1c/1b(600) (SEQ ID NO: 38) having a low repetitive nucleotide sequence encoding a proline/alanine-rich sequence according to this invention. Double-stranded plasmid DNA was sequenced using the primer XLP-1 (SEQ ID NO: 3), which specifically hybridizes within the coding region of the PAS#1b(200) nucleotide sequence unit (SEQ ID NO: 19). In this electropherogram more than 900 bases are readable and correspond to the known nucleotide sequence of pXL2-PAS#1f/1c/1b(600). Notably, the electropherogram shows no signs of unspecific or multiple primer hybridization.

Figure 4: pASK75-PAS#1f/1c/1b(600)-IL1Ra, a genetically stable expression vector for bacterial production of the PAS#1(600)-IL1Ra fusion protein having therapeutic relevance.

(A) Plasmid map of pASK75-IL1Ra (SEQ ID NO: 49) and (B) of its derivative pASK75-PAS#1f/1c/1b(600)-IL1Ra (SEQ ID NO: 50) after insertion of a PAS#1f/1c/1b(600) gene cassette. The structural gene for the biologically/pharmacologically active (pre)protein PAS#1(600)-IL1Ra comprising the low repetitive nucleotide sequence encoding a PAS#1 polypeptide with 601 amino acid residues and the structural gene for human IL-1Ra as well as coding regions for the bacterial OmpA signal sequence and a His₆-tag is cloned under transcriptional control of the *tet* promoter/operator ($tet^{p/o}$). The plasmid backbone outside the expression cassette flanked by the *Xba*I and *Hind*III restriction sites is identical with that of the generic expression vector pASK75 (Skerra (1994) Gene 151:131-135). The singular *Sap*I restriction site that was retained after insertion of the low repetitive nucleic acid molecule

- 90 -

encoding a proline/alanine-rich amino acid repeat sequence according to this invention is indicated.

Figure 5: Analysis of the genetic stability of the low repetitive nucleic acid molecule encoding a proline/alanine-rich amino acid repeat sequence, PAS#1f/1c/1b(600), in comparison with the prior art nucleotide sequence PAS#1a(600)

Agarose gel electrophoresis after *XbaI/HindIII* restriction analysis of 10 plasmid preparations of *E. coli* JM83 clones transformed with pASK75-PAS#1a(600)-IL1Ra (lanes 1-5) (SEQ ID NO: 51) or pASK75-PAS#1f/1c/1b(600)-IL1Ra (lanes 1-10) (SEQ ID NO: 50) cultivated over 7 days, which corresponds to approximately 70 generations of bacterial cell division. 4 of 5 analyzed clones of pASK75-PAS#1a(600)-IL1Ra revealed shortened DNA fragments comprising the nucleic acid insert encoding the proline/alanine-rich amino acid repeat sequence (Fig. 5A), clearly indicating genetic instability. In contrast, all 5 clones of pASK75-PAS#1f/1c/1b(600)-IL1Ra showed only the expected bands corresponding to 3093 bp and 2377 bp, respectively, indicating an intact nucleic acid insert encoding the proline/alanine-rich amino acid repeat sequences and high genetic plasmid stability. Thus, low repetitive nucleotide sequences encoding proline/alanine-rich amino acid repeat sequences according to this invention offer a clear advantage over the repetitive nucleotide sequences of the prior art.

Figure 6: Seamless cloning of a low repetitive nucleotide sequence encoding proline/alanine-rich amino acid repeat sequences on an expression plasmid encoding the biologically active protein IL-1Ra.

(A) Schematic illustration of a coding region for a fusion protein comprising the OmpA signal sequence gene followed by a GCC alanine codon, a first *SapI* recognition sequence GCTCTTC on the non-coding strand, a GC dinucleotide spacer, and a second *SapI* recognition in reverse orientation, with its recognition sequence GCTCTTC on the coding strand, finally followed by a GCC alanine codon and the coding sequence for mature IL-1Ra (UniProt ID P18510). The entire sequence shown was cloned via *XbaI/HindIII* restriction sites on the generic expression vector pASK75. (B) Schematic illustration of the DNA cassette described in (A) after *SapI* cleavage and excision of the short 24 bp insert flanked by the two *SapI* restriction sites. Of note, due to the two 5'-overhangs only the middle 18 nucleotides form a DNA double strand and thus comprise base pairs in the true sense. (C) Seamless insertion in a unidirectional manner of the DNA fragment comprising the low repetitive nucleotide sequence PA#1b(200), e.g., excised from pXL2-PA#1b(200) (SEQ ID NO: 54), via compatible sticky GCC/CGG ends generated by *EarI* restriction digest (cf. Figure 1). The resulting expression cassette comprising the low repetitive nucleotide sequence encoding a proline/alanine-rich amino acid repeat

sequence according to this invention is shown as SEQ ID NO: 47 herein further below.

Figure 7: Characterization of the PAS#1(600)-IL1Ra fusion protein produced in *E. coli* using the genetically stable expression vector pASK75-PAS#1f/1c/1b(600)-IL1Ra

(A) Analysis of the PAS#1(600)-IL1Ra fusion protein purified via IMAC, AEX, CEX and SEC by 10 % SDS-PAGE. The gel shows 2 µg samples of PAS#1(600)-IL1Ra reduced with 2-mercaptoethanol (lane 1) and also not reduced (lane 2). Sizes of the protein marker proteins (M) are indicated on the left. The PAS#1(600)-IL1Ra fusion protein appears as a single homogeneous band with an apparent molecular size above 116 kDa. Due to poor SDS binding, PAS fusion proteins generally show significantly larger sizes (Schlapschy (2013) Protein Eng Des Sel. 26:489-501) than, e.g., the calculated mass of 68 kDa for PAS#1(600)-IL1Ra. (B) Characterization of the PAS#1(600) fusion protein via Electrospray Ionisation Mass Spectrometry (ESI-MS). A deconvoluted ESI-MS spectrum of the purified PAS#1(600)-IL1Ra fusion protein reveals a measured mass of 67994.8 Da, which almost perfectly corresponds to the calculated mass of 67994.9 Da.

Figure 8: Characterization of a pure PA#1(600) polypeptide produced in *E. coli* using the genetically stable expression vector pASK37-MP-PA#1d/1c/1b(600)

(A) Plasmid map of pASK37-MP-PA#1d/1c/1b(600) (SEQ ID NO: 53). The low repetitive nucleotide sequence encoding the PA#1(600) polypeptide was cloned under transcriptional control of the lacUV5 promoter/operator (lacUV5^{P/O}) preceded by codons for a start Met residue and a Pro residue. The plasmid backbone outside the expression cassette flanked by the *Xba*I and *Hind*III restriction sites is identical with that of the generic expression vector pASK37 (Skerra (1991) Protein Eng. 4:971-979). (B) Analysis of the recombinant PA#1(600) polypeptide by 10 % SDS-PAGE, followed by staining with barium iodide. Loaded samples are lysed cells after 3 h expression (lane 1), protein precipitated with 20 % w/v (NH₄)₂SO₄ (lane 2), the supernatant of the (NH₄)₂SO₄ precipitate dissolved in 20 mM Tris/HCl after centrifugation for 20 min at 17.000 rpm and a subsequent treatment with 1 % v/v AcOH (lane 3) and the flow through of subsequent subtractive cation exchange chromatography of the PA#1(600) polypeptide treatment (lane 4). The PA#1(600) polypeptide poorly binds SDS; thus, the PA#1(600) polypeptide appears as a red/yellow iodine-stained band within the stacking gel, indicating homogeneous composition. (C) The gel shown in (B) after destaining with water and staining with Coomassie brilliant blue. The PA#1(600) polypeptide poorly binds SDS and does not stain by the Coomassie dye; thus, only impurities (host cell proteins) are visible on the Coomassie stained gel. (D) Characterization of the pure PA#1(600) polypeptide via Electrospray Ionisation Mass Spectrometry (ESI-MS). The deconvoluted spectrum reveals a measured

- 92 -

mass of 48301.78 Da, which almost perfectly matches the calculated mass for the recombinant PA#1(600) polypeptide, carrying an additional Pro residue at the N-terminus as explained above and an additional Ala residue at the C-terminus due to the *SapI* restriction site that was employed for gene cloning (48301.4 Da). Note that this recombinant polypeptide no longer carries the start Met residue, most likely resulting from the intracellular action of methionine aminopeptidase (Gigliore (2015) *Biochimie* 114:134-46).

Figure 9: Automated repeat analysis of nucleotide sequences encoding proline/alanine-rich amino acid sequences.

Repeats of natural as well as synthetic nucleotide sequences encoding proline/alanine-rich amino acid sequences of the prior art were analyzed as described in Example 13 using the Python script NRS-Calculator (see Example 14) and compared here to the low repetitive nucleotide sequences PAS#1b(200) and PA#1e/1d/1c/1b(800) according to this invention. The frequency (number of occurrences) of all repeats of a certain length within the analyzed nucleotide sequence was plotted against the repeat length. (A) Histogram of the prior art nucleotide sequence PAS#1a(200) (SEQ ID NO: 11). (B) Histogram of the nucleotide sequence PA#1a(200) (SEQ ID NO: 14) (C) Histogram of the nucleotide sequence encoding the glycomodule [(AlaPro)₅]₂₀APA (SEQ ID NO: 16). (D) Histogram of the nucleotide sequence encoding the glycomodule [AAPAPAPAP]₁₀AS (SEQ ID NO: 17). (E) Histogram of the nucleotide sequence encoding a proline/alanine-rich stretch within the large tegument protein of the macacine herpesvirus 1 (SEQ ID NO: 18). (F) Histogram of the low repetitive nucleotide sequence PAS#1b(200) (SEQ ID NO: 19). (G) Close-up view of the PAS#1b(200) data illustrated in (F). (H) Histogram of the low repetitive nucleotide sequence PA#1e/1d/1c/1b(800) according to this invention (SEQ ID NO: 44). (I) Close-up view of the PA#1e/1d/1c/1b(800) data illustrated in (H). The total length of the analyzed nucleotide sequence (N_{tot}) and its Nucleotide Repeat Score (NRS), which is a measure to assess the quality of nucleic acid molecules encoding proline/alanine-rich amino acid sequences with regard to the frequency and lengths of repeats, are summarized for the different nucleotide sequences in Tables 1 and 2.

Figure 10: pASK75-PA#1d/1c/1b(600)-IL1Ra, a genetically stable expression vector for bacterial production of the PA#1(600)-IL1Ra fusion protein having therapeutic relevance.

Plasmid map of pASK75-PA#1d/1c/1b(600)-IL1Ra (SEQ ID NO: 77). The structural gene for the biologically/pharmacologically active (pre)protein PA#1(600)-IL1Ra comprising the low repetitive nucleotide sequence encoding a PA#1 polypeptide with 601 amino acid residues and the structural gene for human IL-1Ra as well as coding regions for the bacterial OmpA signal

- 93 -

sequence and a His6-tag is cloned under transcriptional control of the *tet* promoter/operator ($tet^{p/o}$). The plasmid backbone outside the expression cassette flanked by the *Xba*I and *Hind*III restriction sites is identical with that of the generic expression vector pASK75 (Skerra (1994) Gene 151:131-135). The singular *Sap*I restriction site that was retained after insertion of the low repetitive nucleic acid molecule encoding a proline/alanine-rich amino acid repeat sequence according to this invention is indicated.

Figure 11: Analysis of genetic stability of the low repetitive nucleic acid molecule PA#1d/1c/1b(600), encoding a proline/alanine-rich amino acid repeat sequence, in comparison with the prior art nucleotide sequence PA#1a(600)

Agarose gel electrophoresis after *Xba*I/*Hind*III restriction analysis of 10 plasmid preparations of *E. coli* JM83 clones transformed with pASK75-PA#1d/1c/1b(600)-IL1Ra (**Fig. 10**) (SEQ ID NO: 77) or 10 plasmid preparations of pASK75-PA#1a(600)-IL1Ra (SEQ ID NO: 78) cultivated over 7 days, which corresponds to approximately 70 generations of bacterial cell division. Lanes: M, molecular size standard (GeneRuler 1 kb DNA Ladder: 500, 750, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 5000, 6000, 8000 and 10000 bp); 1 to 10: plasmid samples from individual clones after restriction digest. At least 4 of 10 analyzed clones of pASK75-PA#1a(600)-IL1Ra revealed shortened DNA fragments comprising the nucleic acid insert encoding the proline/alanine-rich amino acid repeat sequence (**Fig. 11 A**), clearly indicating genetic instability. In contrast, all 10 clones of pASK75-PA#1d/1c/1b(600)-IL1Ra (**Fig. 11 B**) showed only the expected bands corresponding to 3093 bp and 2377 bp, respectively, indicating an intact nucleic acid insert encoding the proline/alanine-rich amino acid repeat sequences and high genetic plasmid stability. Thus, low repetitive nucleotide sequences encoding proline/alanine-rich amino acid repeat sequences according to this invention offer a clear advantage over the repetitive nucleotide sequences of the prior art.

Figure 12: Construction of genetically stable expression vectors for bacterial production of human Leptin fused with proline/alanine-rich amino acid sequences

(A) Plasmid map of pASK37-MP-huLeptin (SEQ ID NO: 81) containing a seamless cloning cassette flanked by *Sap*I restriction sites to allow seamless and directed cloning of a low repetitive nucleotide sequence encoding proline/alanine-rich amino acid repeat sequences in frame with the structural gene of human Leptin. (B) Plasmid map of pASK37-MP-huLeptin-PA#1d/1c/1b(600) (SEQ ID NO: 82), a derivative of pASK37-MP-huLeptin with the insertion of a PA#1d/1c/1b(600) gene cassette (SEQ ID NO: 42). (C) Plasmid map of pASK37-MP-huLeptin-PAS#1f/1c/1b(600) (SEQ ID NO: 83), a derivative of pASK37-MP-huLeptin with the insertion of a PAS#1f/1c/1b(600) gene cassette (SEQ ID NO: 38). The structural genes for the biologically/pharmacologically active human protein Leptin, the human Leptin fused to the low

- 94 -

repetitive nucleotide sequence encoding the PA#1(600) and the human Leptin fused to the low repetitive nucleotide sequence encoding the PAS#1(600) polypeptide were cloned under transcriptional control of the *lacUV5* promoter/operator (*lacUV5^{p/o}*), all preceded by codons for a start Met residue and a Pro residue. The plasmid backbone outside the expression cassette flanked by the *Xba*I and *Hind*III restriction sites is identical with that of the generic expression vector pASK37 (Skerra (1991) Protein Eng. 4:971-979).

Figure 13: Characterization of a human Leptin variant fused with a PA#1(600) polypeptide and produced in *E. coli* using the genetically stable expression vector pASK37-MP-PA#1d/1c/1b(600)-huLeptin(W100Q)

(A) SDS-PAGE analysis of the PA#1(600)-huLeptin(W100Q) fusion protein using a 10 % polyacrylamide gel followed by staining with Coomassie brilliant blue R-250. The gel shows a protein molecular weight (MW) marker (lane M; Thermo Fisher Scientific, Waltham, MA), *E. coli* whole cell extract after 19 h expression under reducing sample conditions (lane 1), protein precipitated with 1 M (NH₄)₂SO₄ reduced (lane 2) and not reduced (lane 5), the protein after anion exchange chromatography reduced (lane 3) and not reduced (lane 6), and the protein after size exclusion chromatography reduced (lane 4) and not reduced (lane 7). PA#1(600)-huLeptin(W100Q) appears as a single band, indicating homogeneous composition. **(B)** Characterization of the PA#1(600)-huLeptin(W100Q) fusion protein via Electrospray Ionisation Mass Spectrometry (ESI-MS). The deconvoluted spectrum reveals a measured mass of 64249.5 Da, which matches the calculated mass for the recombinant fusion protein (64249.8 Da), indicating the successful cleavage of the start Met residue by the bacterial methionine aminopeptidase.

The present invention is additionally described by way of the following illustrative non-limiting examples that provide a better understanding of the present invention and of its many advantages.

Example 1: Synthesis of low repetitive nucleotide sequence units encoding proline/alanine-rich amino acid repeat sequences

A set of different nucleotide sequences, each encoding a proline/alanine-rich amino acid repeat sequence of 200 residues were optimized, including manual adjustment, with regard to low repetitiveness on the nucleotide level, low GC content, low RNA secondary structure, preferred codon-usage for expression in *E. coli* and avoidance of antiviral motifs as well as CIS-acting elements. To this end, established algorithms such as the condition-specific codon optimization approach (Lanza (2014) BMC Syst Biol 8:33) or the GeneOptimizer algorithm (Raab (2010) Syst Synth Biol 4:215–225) were applied. The initial sequences obtained thereof were manually

adjusted in the following manner.

Repeats longer than a given threshold (e.g., 14 nucleotides) were identified using the Visual Gene Developer software version 1.2, which is freely available at <http://visualgenedeveloper.net>. Subsequently, codons within identified repeats were stepwise substituted. In particular, GC-rich codons within the identified repeats were replaced by AT-rich codons prevalent in highly expressed genes in a host organism of choice (e.g., *E. coli*, *P. pastoris* or CHO). After each substitution, the entire nucleotide sequence was again analyzed for repeats. In case the substitution led to a new repeat longer than the given threshold, the nucleotide exchange(s) was rejected and a different codon within the previously identified repeat was substituted. If this approach failed, two codons within the identified long repeat were substituted in parallel. In this way, all identified repeats above a given threshold were iteratively eliminated while maintaining the encoded proline/alanine-rich amino acid sequence.

In a second step, the codon adaptation index (CAI), GC content and stable mRNA structures of the optimized nucleotide sequence was analyzed using the Visual Gene Developer software and compared to the start sequence. Additional manual adjustments, again by codon substitution / silent mutation, were performed until the optimized nucleotide sequence reached a CAI, GC content or mRNA structure equal or better than the start sequence. The repeat analysis from step 1 was carried out again, and, if necessary, other codons were exchanged in order to meet the objectives, which were repeat threshold, CAI, GC content and mRNA structures (secondary structures).

In a third step, different individually optimized nucleotide sequences, each encoding the same 200-residue proline/alanine-rich amino acid repeat sequence were combined, i.e. appended to each other, and the resulting longer nucleotide sequence was optimized in the same manner as in steps 1 and 2. Finally, the resulting long acid sequence was divided into shorter, e.g., DNA cassettes of 600 nucleotides lengths. For example, the 2400 nucleotide sequence PAS#1d/1f/1c/1b (SEQ ID NO: 39) was divided into four shorter cassettes (SEQ ID NO: 19, 20, 21, 23). Similarly, the 2400 nucleotide sequence PA#1e/1d/1c/1b (SEQ ID NO: 44) was divided into four shorter cassettes (SEQ ID NO: 28, 29, 30, 31), each comprising 600 nucleotides.

Flanked by two *SapI* recognition sites (5'-GCTCTTC-3') in reverse complementary orientation, resulting in 5'-GCC/5'-GGC nucleotide overhangs after restriction enzyme digest, these optimized nucleotide sequence units were individually synthesized by different commercial vendors. Of note, due to the presence of the two GCC/GGC nucleotide overhangs, only the middle 597 nucleotides form a DNA double strand after excision and, hence, comprise base pairs (bp). Also, the optimized 600 nucleotide sequence is extended by an additional Ala codon due to the presence of the second *SapI* restriction site, thus leading to a cloned DNA cassette of

- 96 -

overall 603 nucleotides encoding a proline/alanine-rich amino acid sequence. The presence of the two flanking *SapI* restriction sites enables precise excision and subcloning, e.g., on pXL2, of the entire DNA cassette of the invention.

Further sets of nucleotide sequence units encoding proline/alanine-rich amino acid repeat sequences, codon-optimized for expression in *Escherichia coli*, *Pichia pastoris*, human embryonic kidney (HEK) cells, *Pseudomonas fluorescens*, *Corynebacterium glutamicum*, *Bacillus subtilis*, *Tetrahymena thermophila*, *Saccharomyces cerevisiae*, *Kluyveromyces lactis*, *Physcomitrella patens* or *Cricetulus griseus*, were designed and synthesized in the same manner. Codon preference tables for these organisms are available for download at <http://www.kazusa.or.jp/codon>. The synthesized nucleic acid molecules according to the invention and their nucleotide sequence characteristics are summarized in Table 1.

Example 2: Assembly of low repetitive nucleotide sequence units to longer nucleotide sequences encoding proline/alanine-rich amino acid repeat sequences

Plasmids obtained from commercial vendors, each carrying a cloned synthesized DNA fragment, were digested with *SapI* and the resulting 600 nucleotide DNA fragment was purified via agarose gel electrophoresis according to standard procedures (Sambrook (2001) loc. cit.) The individual nucleotide sequence units were assembled to longer nucleotide sequences using the plasmid pXL2 (SEQ ID NO: 48), a derivative of pUC19 (Yanisch-Perron (1985). Gene. 33, 103-119) shown in Fig. 1A. pXL2 contains a single *SapI* restriction site, and digest with this type IIS restriction enzyme generates a 5'-GCC/5'-GGC-overhang coding for alanine, which is compatible with the sticky ends of the synthesized purified DNA fragments (Fig. 1C). After insertion/ligation of one nucleotide sequence unit, the plasmid can be opened at one end, here upstream, of the cloned sequence unit by another *SapI* restriction digest (Fig. 1D). This vector design allows the stepwise insertion of identical or different low repetitive nucleotide sequence units, yielding longer cloned gene cassettes encoding proline/alanine-rich amino acid repeat sequences (Fig. 1E).

As an example, first the nucleotide sequence unit PAS#1b(200) (SEQ ID NO: 19), then the sequence unit PAS#1c(200) (SEQ ID NO: 20), and subsequently the sequence unit PAS#1f(200) (SEQ ID NO: 23) were inserted into pXL2 via the *SapI* restriction site in the described manner, resulting in the plasmid pXL2-PAS#1f/1c/1b(600) (SEQ ID NO: 38). In a subsequent step, the sequence unit PAS#1d(200) (SEQ ID NO: 19) was additionally inserted in the same manner using the *SapI* restriction site. The resulting plasmid contained the assembled 2400 bp DNA cassette PAS#1d/1f/1c/1b(800) which in total revealed nucleotide sequence repeats with a maximum length of 14 nucleotides (SEQ ID NO: 39). As the recognition sequence of *EarI* (5'-CTCTTC-3') downstream of the low repetitive DNA cassette cloned on

- 97 -

pXL2 is also part of the recognition sequence of *SapI*, the entire assembled DNA cassette can be easily excised via restriction digest with *EarI*, thus cutting twice, allowing subsequent use for further subcloning.

In the same manner, the low repetitive nucleotide sequence PA#1e/1d/1c/1b(800) (SEQ ID NO: 44) was assembled from the nucleotide sequence units PA#1b(200) (SEQ ID NO: 28), PA#1c(200) (SEQ ID NO: 29), PA#1d(200) (SEQ ID NO: 30) and PA#1e(200) (SEQ ID NO: 31) in the stated order. The described assembled nucleotide sequences as well as further exemplary low repetitive nucleic acid molecules encoding proline/alanine-rich amino acid repeat sequences according to this invention, also with codon usage optimized for host organisms different from *E. coli*, are summarized in Table 1. The disclosed cloning strategy offers a simple, stepwise assembly of complex gene cassettes comprising long low repetitive nucleic acid molecules encoding proline/alanine-rich amino acid repeat sequences, which cannot be directly obtained by common gene synthesis methods.

Example 3: Repetitiveness analysis of nucleotide sequences encoding proline/alanine-rich amino acid repeat sequences

A dot plot analysis was performed for different nucleotide sequences encoding the proline/alanine-rich amino acid repeat sequences PA#3 (SEQ ID NO: 15) (**Fig. 2A**) as disclosed in WO 2011144756, PAS#1 (SEQ ID NO: 11) (**Fig. 2B**) as disclosed in WO2008155134, a [(AP)₅]_n multimer (SEQ ID NO: 16) as disclosed in WO2004094590, and a repetitive proline/alanine-rich amino acid sequence region of the very large tegument protein of *Macacine herpesvirus 1* gene, published under the GenBank accession number AAP41454.1 (SEQ ID NO: 18) (**Fig. 2C**). The analysis was performed by aligning each nucleotide sequence to itself using the dot plot tool "dottup" of the Geneious software package version 8.1 (Biomatters, Auckland, New Zealand) and applying a repeat window of 14 or 15 nucleotides. The algorithm of this software is based on the freely available EMBOSS 6.5.7 tool "dottup" (Sanger Institute, Cambridge, UK). Resulting two-dimensional dot plot graphs obtained for the prior art nucleotide sequences were compared to dot plots of the low repetitive nucleotide sequence units PA#3b(200) (SEQ ID NO: 36), PA#1b(200) (SEQ ID NO: 28) and the assembled nucleotide sequences PAS#1f/1c/1b(600) (SEQ ID NO: 38) and PAS#1d/1f/1c/1b(800) (SEQ ID NO: 39) which encode longer proline/alanine-rich amino acid repeat sequences. Whereas all analyzed prior art nucleotide sequences revealed a highly repetitive nature on the nucleotide sequence level, as illustrated by black diagonal lines (**Fig. 2 A, B, C**), dot plots of the optimized nucleotide sequences encoding proline/alanine-rich amino acid repeat sequences according to this invention showed only a few scattered or short 14 nucleotide repeats (black lines) within the entire analyzed nucleotide sequence of 600

nucleotides of the PA#3b(200) and PA#1b(200) cassettes (Fig. 2 A, C), the 1800 nucleotide PAS#1f/1c/1b(600) cassette (Fig. 2B) or the 2400 nucleotide PAS#1d/1f/1c/1b(800) cassette (Fig. 2 B).

Example 4: DNA-sequencing of low repetitive nucleic acid molecules encoding long proline/alanine-rich amino acid repeat sequences

The low repetitive PAS#1f/1c/1b(600) DNA cassette (SEQ ID NO: 38) cloned on the plasmid pXL2 and described in Example 2 was sequenced by a DNA-sequencing service provider (Eurofins Genomics, Ebersberg, Germany) using Sanger cycle sequencing on an ABI 3730XL instrument (Thermo Fisher Scientific, Waltham, MA). To this end, 8 µl (150 ng/µl) of pXL2-PAS#1f/1c/1b(600) plasmid DNA, isolated from transformed *E. coli* XL1-blue cells using the QIAprep Spin Miniprep kit (Qiagen, Hilden, Germany) was mixed with 5 µl doubly distilled H₂O and 2 µl primer XLP-1 (10 µM) (SEQ ID NO: 3), which hybridizes within the coding region of the PAS#1b(200) nucleotide sequence unit and submitted to the DNA-sequencing service provider. As result, an error-free electropherogram comprising more than 900 assignable nucleotides (Fig. 3) was obtained, which showed no signs of unspecific or multiple primer binding. Thus, in contrast to long repetitive nucleotide sequences, which can only be partially sequenced using primers hybridizing to vector nucleotide sequences upstream or downstream of the cloned DNA, long low repetitive nucleic acid molecules according to this invention encoding proline/alanine-rich amino acid repeat sequences can be easily sequenced also using primers that specifically hybridize internally, within the cloned nucleotide sequence. This enables multiple overlapping sequence reads using different suitable primers, thus allowing full sequence coverage even of very long nucleic acid molecules according to the invention.

Example 5: Construction of pASK75-PAS#1f/1c/1b(600), a genetically stable expression vector for the bacterial production of a therapeutic PAS#1(600)-IL1Ra fusion protein

For the construction of an expression plasmid encoding the interleukin-1 receptor antagonist (IL-1Ra) as fusion with a 600 residue PAS#1 amino acid repeat sequence (SEQ ID NO: 38), the vector pASK75-IL1Ra (Fig. 4 A) (SEQ ID NO: 49) was cut with *SapI*, dephosphorylated with shrimp alkaline phosphatase (Thermo Fisher Scientific, Waltham, MA) and ligated with a DNA fragment corresponding to the low repetitive nucleotide sequence cassette encoding the 600 residue PAS#1 polypeptide, which was excised from the plasmid pXL2-PAS#1f/1c/1b(600) by restriction digest with *EcoRI*. After transformation of *E. coli* XL1-Blue (Bullock (1987) Biotechniques 5:376-378), plasmid DNA was prepared and the presence of the inserted DNA fragment was confirmed by restriction analysis and DNA sequencing. The

- 99 -

resulting plasmid was designated pASK75-PAS#1f/1c/1b(600)-IL1Ra (SEQ ID NO: 50) and is shown in **Fig. 4 B**.

Example 6: Long-term genetic stability testing of a plasmid harboring a low repetitive nucleic acid molecule encoding a proline/alanine-rich amino acid repeat sequence

The genetic stability of the plasmid pASK75-PAS#1f/1c/1b(600)-IL1Ra (SEQ ID NO: 50) was compared to the genetic stability of pASK75-PAS#1a(600)-IL1Ra (SEQ ID NO: 51), a derivative wherein the PAS#1f/1c/1b(600) DNA cassette was substituted by the repetitive nucleic acid PAS#1a(600) (SEQ ID NO: 12). To this end, *E. coli* KS272 (Strauch (1988) Proc. Natl. Acad. Sci. USA 85:1576-1580) was transformed with the respective plasmid using the calcium chloride method (Sambrook (2001) loc. cit.) and cultured for 7 days at 37 °C, 170 rpm, in 50 ml Luria Bertani (LB) medium supplemented with 100 mg/mL ampicillin in a 100 mL shake flask without induction of gene expression. During this period, bacterial cells were twice daily (in the morning and in the evening) transferred into fresh medium using a 1:1000 dilution. On day 7, after a continuous growth over approximately 70 generations, the culture was finally grown to stationary phase and cells were plated on LB/Amp agar. Then, individual clones were picked, used for inoculation of 50 mL cultures in LB medium and, after growth to stationary phase over night, plasmid DNA from five clones for each of the two plasmids was prepared using the Qiagen Miniprep Kit (Qiagen, Hilden, Germany) and analyzed by a *Xba*I/*Hind*III restriction digest (**Fig. 5**).

Only 1 out of 5 analyzed clones of pASK75-PAS#1a(600)-IL1Ra showed the expected bands corresponding to 3093 bp and 2377 bp (**Fig. 5, lane 1**). Two clones (**Fig. 5, lanes 3 and 5**) revealed a band at 573 bp, the approximate size of the combined gene sequences encoding OmpA and IL1Ra, indicating a more or less complete loss of the repetitive PAS#1a(600) sequence cassette, possibly by recombination. Two other clones showed significantly shortened DNA fragments (**Fig. 5, lanes 2 and 4**), also indicating deletion events within the repetitive PAS#1a(600) sequence cassette and, hence, genetic instability. In contrast, all five analyzed clones of pASK75-PAS#1f/1c/1b(600)-IL1Ra revealed the expected bands at 3093 bp and 2377 bp (**Fig. 5, lanes 6–10**), indicating an intact gene cassette encoding proline/alanine-rich amino acid repeat sequences and, thus, high genetic plasmid stability of the low repetitive nucleic acid molecules according to this invention.

Example 7: Seamless and directed cloning of a low repetitive nucleotide sequence

encoding proline/alanine-rich amino acid repeat sequences on an expression plasmid encoding the biologically active protein IL-1Ra.

With the goal of pharmaceutical application, fusion proteins comprising solely the biologically active protein and a proline/alanine-rich amino acid repeat sequence are desired. The absence of additional amino acid linkers, e.g., introduced in order to provide or utilize restriction sites for cloning, may prevent potential immune responses during clinical use and/or avoid unintended interactions on the protein level. Therefore, a seamless cloning strategy was developed (**Fig. 6**) for the directed insertion of low repetitive nucleotide sequences, here exemplified for the DNA fragment comprising PA#1b(200) (SEQ ID NO: 28), on a derivative of the generic expression plasmid pASK75 (Skerra (1994) loc. cit.) encoding the biologically active protein IL1-Ra (Molto (2010) Joint Bone Spine. 77:102-107).

At first, a synthetic DNA fragment encoding the mature amino acid sequence of IL1-Ra (UniProt ID P18510) was obtained from a gene synthesis provider (Thermo Fisher Scientific, Regensburg, Germany). This gene fragment (SEQ ID NO: 46) comprised an *Xba*I restriction site, followed by a ribosomal binding site, the nucleotide sequence encoding the OmpA signal peptide, followed by a GCC alanine codon, a first *Sap*I recognition sequence GCTCTTC on the non-coding strand, a GC dinucleotide spacer, and a second *Sap*I restriction sequence in reverse complementary orientation, with its recognition sequence GCTCTTC on the coding strand, followed by a GCC alanine codon directly linked to the coding sequence for mature IL1Ra (UniProt ID P18510), which was finally followed by a *Hind*III restriction site.

This gene fragment was cloned on pASK75 via the flanking restriction sites *Xba*I and *Hind*III according to standard procedures (Sambrook (2001) loc. cit.). The resulting plasmid (cf. **Fig. 6A**) was digested with *Sap*I, which led to the liberation of a small (24 bp) DNA insert containing both *Sap*I recognition sites and a cleaved vector backbone with compatible 5'-GCC/5'-GGC sticky ends at the position directly in front of the encoded mature N-terminus of IL-1Ra, which is ideally suited for insertion of the low repetitive nucleic acid molecule encoding the proline/alanine-rich amino acid repeat sequence (**Fig. 6B**). After isolation of the vector fragment using the QIAquick gel extraction kit (Qiagen, Hilden, Germany) and dephosphorylation with the thermosensitive alkaline phosphatase FastAP (Thermo Fisher Scientific, Waltham, MA), both according to the manufacturer's instructions, it was ligated with the PA#1b(200) gene cassette excised from pXL2-PA#1b(200) (SEQ ID NO: 54) via *Ear*I restriction digest (**Fig. 6C**). The resulting plasmid (SEQ ID NO: 56) allows the bacterial expression of a fusion protein (SEQ ID NO: 10) consisting solely of a proline/alanine-rich amino acid repeat sequence fused with the biologically active protein IL-1Ra (after *in vivo* processing of the OmpA signal peptide upon periplasmic secretion in *E. coli*).

Example 8: Bacterial production and purification of a fusion protein between the

PAS#1(600) sequence and IL-1Ra encoded on the genetically stable plasmid pASK75-PAS#1f/1c/1b(600)-IL1Ra

The PAS#1(600)-IL1-Ra fusion protein (calculated mass: 68 kDa) was produced at 25 °C in *E. coli* KS272 harboring the genetically stable expression plasmid pASK75-PAS#1f/1c/1b(600)-IL1Ra from Example 6 and the folding helper plasmid pTUM4 (Schlupschy (2006) Protein Eng. Des. Sel. 20:273-284) using an 8 L bench top fermenter with a synthetic glucose mineral medium supplemented with 100 mg/L ampicillin and 30 mg/L chloramphenicol according to a published procedure (Schiweck (1995) Proteins 23:561-565). Recombinant gene expression was induced by addition of 500 µg/L anhydrotetracycline (Skerra (1994) loc. cit.) as soon as the culture reached OD₅₅₀ = 28. After an induction period of 2.5 h, cells were harvested by centrifugation and resuspended during 10 min in ice-cold periplasmic fractionation buffer (500 mM sucrose, 1 mM EDTA, 100 mM Tris/HCl pH 8.0; 2 ml per L and OD₅₅₀). After adding 15 mM EDTA and 250 µg/mL lysozyme, the cell suspension was incubated for 20 min on ice, centrifuged several times, and the cleared supernatant containing the recombinant protein was recovered.

The periplasmic extract was dialyzed four times at 4 °C against 5 L 40 mM Na-phosphate pH 7.5, 500 mM NaCl, respectively and purified by means of the His₆-tag using an 80 ml HisTrap HP column (GE Healthcare, Freiburg, Germany). The protein was eluted with an imidazole/HCl pH 7.5 concentration gradient from 0 to 200 mM in 40 mM Na-phosphate pH 7.5, 0.5 M NaCl. The purified protein was pooled and dialyzed twice against 5 L 20 mM Tris/HCl pH 8.0, 1 mM EDTA at 4 °C for at least 6 h, respectively. The dialyzed protein solution was subjected to anion exchange chromatography using a 60 ml XK column (GE Healthcare, Freiburg, Germany) packed with Source15Q resin, connected to an Äkta purifier system (GE Healthcare, Freiburg, Germany), using 20 mM Tris/HCl pH 8.0, 1 mM EDTA as running buffer. The protein was eluted using a NaCl concentration gradient from 0 to 200 mM in running buffer.

Eluted fractions were dialyzed twice against 10 mM MES/HCl pH 6.0, 1 mM EDTA at 4 °C for at least 6 h, respectively, and subsequently subjected to a cation exchange chromatography using an XK column packed with 36 ml Source15S resin (GE Healthcare, Freiburg, Germany). The cation exchange chromatography was performed on an Äkta purifier system using 10 mM MES/HCl pH 6.0, 1 mM EDTA as running buffer and a NaCl concentration gradient from 0 to 500 mM in running buffer over 4 column volumes to elute the protein. The eluted protein fractions containing PAS#1(600)-IL1-Ra were again pooled, dialyzed against 5 L phosphate-buffered saline (PBS: 115 mM NaCl, 4 mM KH₂PO₄ and 16 mM Na₂HPO₄ pH 7.4) at 4 °C overnight, concentrated to 5 mg/ml using an Amicon Ultra centrifugal filter device (30000 MWCO; 15 mL; Millipore, Billerica, MA) and further purified via size exclusion chromatography using a HiLoad 26/60 Superdex 200 prepgrade column (GE Healthcare,

Freiburg, Germany) equilibrated with PBS.

A homogeneous protein preparation without signs of aggregation was obtained with a final yield of 70 mg from one 8 L fermenter. Protein concentration was determined by measuring the absorption at 280 nm using a calculated extinction coefficient (Gill (1989) Anal. Biochem. 182:319-326) of $15720 \text{ M}^{-1} \text{ cm}^{-1}$. SDS-PAGE was performed using a high molarity Tris buffer system (Fling (1986) Anal. Biochem. 155:83-88) (**Fig. 7A**).

Example 9: ESI-MS analysis of the PAS#1(600)-IL1Ra fusion protein

PAS#1(600)-IL1Ra produced and purified as described in Example 8 was dialyzed twice against a 1000-fold volume of 10 mM ammonium acetate pH 6.8 and analyzed via ESI mass spectrometry on a Q-ToF Ultima instrument (Waters, Eschbronn, Germany) using the positive ion mode. The deconvoluted spectrum of the PA#1(600)-IL1Ra fusion protein revealed a mass of 67994.8 Da, which essentially coincides with the calculated mass of 67994.8 Da (**Fig. 7B**). This clearly demonstrates that the entire PA#1(600)-IL1Ra fusion protein can be efficiently produced in *E. coli* using the genetically stable expression plasmid pASK75-PAS#1f/1c/1b(600)-IL1Ra.

Example 10: Construction of pASK37-MP-PA#1d/1c/1b(600), a genetically stable plasmid for the production of a proline/alanine-rich amino acid repeat polypeptide in *E. coli*

For the construction of a stable expression plasmid encoding the pure PA#1(600) polypeptide, 100 pmol of the primers NdeI-MP-SapI-HindIII_{fw} (SEQ ID NO: 4) and NdeI-MP-SapI-HindIII_{rev} (SEQ ID NO: 5) were phosphorylated, mixed, heated up to 80 °C for 10 min and slowly cooled down to room temperature overnight to allow hybridization. The resulting double stranded DNA fragment exhibited sticky ends compatible to *NdeI* and *HindIII* overhangs. The plasmid pASK37 (Skerra (1991) loc. cit) was cut with *NdeI* and *HindIII* and the backbone fragment was ligated with the hybridized primers.

The resulting plasmid was digested with *SapI*, which led to the liberation of a small (24 bp) insert containing two *SapI* recognition sites and a cleaved vector backbone with compatible sticky 5'-GCC/5'-GGC ends. These sticky ends are ideally suited for insertion of the low repetitive nucleotide sequence encoding the proline/alanine-rich amino acid repeat sequence at the position directly downstream of the N-terminal start methionine codon (ATG) followed by the proline codon CCA, which was found to allow efficient translational initiation. After isolation of the vector fragment using the QIAquick gel extraction kit and dephosphorylation with the thermosensitive alkaline phosphatase FastAP according to the manufacturer's instructions, it was ligated with the low repetitive gene cassette PA#1d/1c/1b(600) (SEQ ID

NO: 42) excised from pXL2-PA#1d/1c/1b(600) via *EarI* restriction digest. The resulting plasmid (SEQ ID NO: 53) permits expression of a polypeptide comprising solely a proline/alanine-rich amino acid repeat sequence (Fig. 8A).

Example 11: Bacterial expression and purification of a PA#1(600) polypeptide encoded on the genetically stable plasmid pASK37-MP-PA#1d/1c/1b(600)

The PA#1(600) polypeptide, with an additional Pro residue at the N-terminus and an additional Ala residue at the C-terminus (calculated mass: 48302 Da), was produced in the cytoplasm of *E. coli* KS272 harboring the expression plasmid pASK37-PA#1d/1c/1b(600) described in Example 10. 4 ml LB medium in a sterile 13 mL polypropylene tube (Sarstedt, Nümbrecht, Germany), substituted with 1 % w/v glucose and 100 mg/L ampicillin, were inoculated with a colony of *E. coli* KS272 transformed with pASK37-PA#1d/1c/1b(600) and grown overnight at 37 °C, 170 rpm. Bacterial protein production was performed at 30 °C in a 5 L shake flask with 2 L terrific broth (TB) medium (Sambrook (2001) loc. cit.) supplemented with 2.5 g/L D-glucose and 100 mg/L ampicillin.

E. coli cultures were inoculated with 2 ml overnight culture, cells were grown overnight and recombinant gene expression was induced at $OD_{550} = 5$ by addition of isopropyl- β -D-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM. Bacteria were harvested 3 h after induction, resuspended in 20 ml 40 mM Na-phosphate pH 7.2, 1 mM EDTA and lysed using a French pressure cell (Thermo Scientific, Waltham, MA). After centrifugation (17,000 rpm, 1 h, 4 °C) of the lysate, no inclusion bodies were observed. The supernatant containing the soluble PA#1(600) polypeptide was subjected to an ammonium sulfate precipitation by stepwise addition of solid $(NH_4)_2SO_4$ to a final concentration of 20 % w/v under continuous stirring at room temperature. The supernatant was centrifuged at 17,000 rpm at room temperature for 20 min. The sediment containing the precipitated PA#1(600) polypeptide was dissolved in 20 mM Tris/HCl pH 8.0 and the solution was centrifuged (13,000 rpm, 10 min, room temperature) to remove insoluble contaminants.

Pure acetic acid (Sigma-Aldrich, Steinheim, Germany) was added to a final concentration of 1 % v/v and impurities were sedimented by centrifugation at 13,000 rpm for 10 min. The supernatant containing the almost pure PA#1(600) polypeptide was dialyzed against a 100-fold volume of 1 % v/v acetic acid overnight at 4 °C. To remove residual impurities, the dialysed protein was subjected to a subtractive cation exchange chromatography using a 1 ml Source15S column (GE Healthcare, Freiburg, Germany) connected to an Äkta purifier system using 1 % v/v acetic acid as running buffer.

Samples from each purification step were analyzed by SDS-PAGE using a high molarity Tris buffer system (Fling (1986) loc. cit.). After SDS-PAGE, the gel was first stained with barium

iodide as described for the analysis of PEG (Kurfurst (1992) Anal. Biochem. 200:244-248). Briefly, the polyacrylamide gel was rinsed with water and then incubated in a 2.5 % w/v BaI₂ (barium iodide dihydrate; Sigma-Aldrich, Steinheim, Germany) solution in water for 5 min. After rinsing with water, the gel was transferred into Lugol solution (10 % w/v p.a. grade KI (AppliChem, Darmstadt, Germany) 5 % p.a. grade I₂ (Riedel de Haen AG, Seelze, Germany) in water) for 5 min. After destaining in 10 % v/v acetic acid, orange PA#1(600) polypeptide bands became visible (**Fig. 8B**). Subsequently, the gel was destained with water and subjected to a second staining with Coomassie brilliant blue R250 (Applichem), dissolved in 10 % acetic acid (Honeywell Specialty Chemicals, Seelze, Germany), 65 % H₂O and 25 % isopropanol (CLN, Niederhummel, Germany). After destaining in 10 % v/v acetic acid blue protein bands (for host cell proteins) became visible (**Fig. 8C**).

Example 12: ESI-MS analysis of a pure PA#1(600) polypeptide

200 µl of the isolated PA#1(600) polypeptide from Example 11 at a concentration of 5 mg/mL was applied to a 1 mL Resource RPC column (GE Healthcare, Freiburg, Germany) connected to an Äkta purifier system using 2 % v/v acetonitrile, 1 % v/v formic acid as running buffer. The protein was eluted using an acetonitrile gradient from 2 % v/v acetonitrile, 1 % v/v formic acid to 80 % v/v acetonitrile, 0.1 % v/v formic acid over 20 column volumes. The eluted protein was directly analyzed via ESI mass spectrometry on a Q-ToF Ultima instrument using the positive ion mode. The deconvoluted spectrum of the PA#1(600) polypeptide revealed a mass of 48301.78 Da, which essentially coincides with the calculated mass of the PA#1(600) polypeptide, with an additional Pro residue at the N-terminus and an additional Ala residue at the C-terminus but devoid of the start methionine (48301.4 Da) (**Fig. 8D**). This clearly demonstrates that a pure PA#1(600) polypeptide (without an affinity tag), encoded by a genetically stable nucleotide sequence, can be produced in *E. coli* in its intact form.

Example 13: Repeat analysis of nucleotide sequences encoding proline/alanine-rich amino acid sequences

As a measure to assess the quality of nucleic acid molecules encoding proline/alanine-rich sequences with regard to the frequency (occurrence) of nucleotide sequence repeats we have devised the Nucleotide Repeat Score (NRS), which is calculated according to the following formula:

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{i=1}^{k(n)} f_i(n)}}{N_{tot}}$$

- 105 -

In this formula, N_{tot} is the total length of the nucleotide sequence analyzed, n is the length of a sequence repeat within the nucleotide sequence analyzed and the frequency $f_i(n)$ is the number of occurrences of this sequence repeat. In case there are several different sequence repeats with the same length n , these different sequence repeats are distinguished by the index i and the number of different sequence repeats with the same length n is $k(n)$. If there is just one type of sequence repeat with length n , $k(n)$ equals 1. The NRS is defined as the sum of the squared repeat length multiplied with the root of the respective overall frequency, divided through the total length of the analyzed nucleotide sequence. The minimal repeat length considered for the calculation of NRS comprises 4 nucleotides, which includes all nucleotide sequences longer than one codon triplet, and it ranges up to $N_{\text{tot}} - 1$, that is the length of the longest nucleotide sequence repeat that can occur more than once in the analyzed nucleotide sequence.

In this context the term repeat means that a nucleotide sequence occurs at least twice within the nucleotide sequence analyzed. When counting the frequencies we have considered both nucleotide stretches with identical sequence that occur at least twice as well as different sequences of the same length which each also occur at least twice. For example, if the overall frequency of a 14mer repeat is five, this can mean either that the same 14mer nucleotide stretch occurs 5 times, or one 14mer nucleotide sequence occurs twice and a different 14 nucleotide sequence occurs three times in the analyzed nucleotide sequence.

Furthermore, each shorter repeat contained within a longer nucleotide sequence repeat is counted separately. For example, if the analyzed nucleotide sequence contains two GCACC nucleotide stretches (i.e., repeats), GCAC and CACC repeats are also counted individually, regardless if they occur within said GCACC nucleotide stretch or, possibly, in addition elsewhere within the analyzed nucleotide sequence. Of note, only repeats on the coding strand of the nucleic acid molecule are considered.

A person skilled in the art can identify nucleotide sequence repeats either manually or with the aid of generic software programs such as the Visual Gene Developer (Jung (2011) loc. cit.), available for download at <http://www.visualgenedeveloper.net>, or the Repfind tool (Betley (2002) loc. cit), available at <http://zlab.bu.edu/repfind>. However, not every algorithm detects each kind of repeat, e.g., the result of the Visual Gene Developer does not include overlapping repeats. Thus, results of software tools have to be checked and, if necessary, manually corrected. Alternatively, the algorithm termed NRS-Calculator described in Example 14 can be used to unambiguously identify nucleotide sequence repeats and to calculate the NRS automatically.

Natural as well as certain synthetic nucleic acids encoding proline/alanine-rich amino acid sequences are known in the art. However, all those sequences are highly repetitive on the genetic level as it becomes clearly evident from the NRS analysis described below and, thus,

their use for biotechnological and/or biopharmaceutical applications is limited.

Several prior art nucleotide sequences encoding proline/alanine-rich amino acid sequences were compared to low repetitive nucleic acid molecules encoding proline/alanine-rich amino acid repeat sequences according to this invention using the NRS-Calculator described in Example 14: the nucleotide sequence PAS#1a(200) (SEQ ID NO: 11) disclosed in WO 2008/155134 (**Fig. 9A**), the nucleotide sequence PA#1a(200) (SEQ ID NO: 14) disclosed in WO2011144756 (**Fig. 9B**), the nucleotide sequence encoding a [(AP)₅]₂₀APA glycomodule (SEQ ID NO: 16) disclosed in US 20060252120 (**Fig. 9C**), the nucleotide sequence of a synthetic gene construct encoding the glycomodule [AAPAPAPAP]₁₀AS (SEQ ID NO: 17) published under GenBank accession number DQ399411.1 (**Fig. 9D**), the 225 nucleotide sequence encoding a proline/alanine-rich sequence within the large tegument protein of the macacine herpesvirus 1 (SEQ ID NO: 18) published under GenBank accession number NP_851896 (**Fig. 9E**), the low repetitive nucleotide sequence PAS#1b(200) (SEQ ID NO: 19) according to this invention (**Fig. 9F,G**) and the low repetitive nucleotide sequence PA#1e/1d/1c/1b(800) (SEQ ID NO: 44) according to this invention (**Fig. 9H,I**).

The calculated repeat frequencies were plotted against the respective repeat length using Kaleidagraph V3.6 software (Synergy Software, Reading, PA) (**Fig. 9**). All histograms of the prior art nucleotide sequences reveal a highly repetitive nature as illustrated by a large number of high bars with broad distribution of repeat lengths, up to very long repeats. Notably, in these cases the repeat frequency decreases only slowly with increasing repeat length (**Fig. 9A-E**). In contrast, the histograms of the low repetitive nucleotide sequences PAS#1b(200) and PA#1e/1d/1c/1b(800) according to this invention show only a few repeats with a maximum length of 14 nucleotides, whose frequencies rapidly decrease to zero when going from shorter to longer repeats (**Fig. 9F,G,H,I**).

The difference in repetitiveness between the prior art nucleotide sequences and the low repetitive nucleotide sequences of the invention becomes even more evident when comparing their Nucleotide Repeat Scores. Whereas all prior art sequences reveal an NRS above 80000 (Table 2), the 600 nucleotide sequence PAS#1b(200) and the 2400 nucleotide sequence PA#1e/1d/1c/1b(800) show NRS values of just 13 and 14, respectively (Table 1). This clearly demonstrates that the repeat quality of the low repetitive nucleotide sequences encoding proline/alanine-rich amino acid repeat sequences according to this invention is much higher compared to prior art sequences, with both fewer and shorter nucleotide sequence repeats.

Table 1: Characteristics of nucleic acid molecules according to this invention

	Low repetitive nucleotide sequence no.	SEQ ID:	Codon-optimized for:	Encoded amino acid repeat	n _{max}	N _{tot}	NRS
A: Nucleotide sequence units (building blocks)							
1	PAS#1b(200)	19	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	13
2	PAS#1c(200)	20	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	12	600	12
3	PAS#1d(200)	21	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	12	600	11
4	PAS#1e(200)	22	<i>CHO</i> (<i>C. griseus</i>)	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	12	600	12
5	PAS#1f(200)	23	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	12	600	11
6	PAS#1g(200)	24	<i>Pichia pastoris</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	24
7	PAS#1h(200)	25	<i>CHO</i> (<i>C. griseus</i>)	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	12	600	20
8	PAS#1i(200)	26	<i>CHO</i> (<i>C. griseus</i>)	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	17
9	PAS#1j(200)	27	<i>CHO</i> (<i>C. griseus</i>)	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	16
10	PA#1b(200)	28	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	600	21
11	PA#1c(200)	29	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	600	18
13	PA#1d(200)	30	<i>E. coli</i>	AAPAAPAPAAP	14	600	19

- 108 -

				AAPAPAAPA (SEQ ID NO: 2)			
14	PA#1e(200)	31	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	600	22
15	PA#1f(200)	32	<i>CHO</i> (<i>C. griseus</i>)	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	600	24
16	PA#1g(200)	33	<i>CHO</i> (<i>C. griseus</i>)	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	600	24
17	PA#1h(200)	34	<i>CHO</i> (<i>C. griseus</i>)	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	32
18	PA#1i(200)	35	<i>CHO</i> (<i>C. griseus</i>)	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	17
19	PA#3b(200)	36	<i>E. coli</i>	AAAPAAAPAAA PAAAPAAAP (SEQ ID NO: 57)	14	600	26
20	PA#5b(198)	37	<i>E. coli</i>	AAAAAPAAAAA PAAAAAP (SEQ ID NO: 58)	14	594	27
101	PA#1j(200)	87	<i>P. pastoris</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	39
102	PA#1k(200)	88	<i>P. pastoris</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	29
103	PA#1l(200)	89	<i>P. pastoris</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	31
104	PA#1m(200)	90	<i>P. pastoris</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	600	24
105	PA#1n(200)	91	<i>S. cerevisiae</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	38
106	PA#1o(200)	92	<i>S. cerevisiae</i>	AAPAAPAPAAP AAPAPAAPA	14	600	20

- 109 -

				(SEQ ID NO: 2)			
107	PA#1p(200)	93	<i>S. cerevisiae</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	14	600	19
108	PA#1q(200)	94	<i>K. lactis</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	600	28
109	PA#1r(200)	95	<i>K. lactis</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	14	600	23
110	PA#1s(200)	96	<i>K. lactis</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	600	34
111	PA#1t(200)	97	<i>H. sapiens</i> (HEK cells)	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	14	600	25
112	PA#1u(200)	98	<i>H. sapiens</i> (HEK cells)	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	600	29
114	PA#1v(200)	99	<i>H. sapiens</i> (HEK cells)	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	600	31
114	PA#1w(200)	100	<i>Bacillus subtilis</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	14	600	23
115	PA#1x(200)	101	<i>Bacillus subtilis</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	16	600	27
116	PA#1y(200)	102	<i>Bacillus subtilis</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	600	32
117	PA#1z(200)	103	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	18	600	45
118	PA#1aa(200)	104	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	14	600	18
119	PA#1ab(200)	105	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	600	25

- 110 -

120	PA#1ac(200)	106	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	600	18
121	PA#1ad(200)	107	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	24
122	PA#1ae(100)	108	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	300	27
123	PA#1af(200)	109	<i>C. glutamicu m</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	600	20
124	PA#1ag(200)	110	<i>C. glutamicu m</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	24
125	PA#1ah(200)	111	<i>C. glutamicu m</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	25
126	PA#1ai(200)	112	<i>C. glutamicu m</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	16	600	21
127	PA#1aj(200)	113	<i>P. patens</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	30
128	PA#1ak(200)	114	<i>P. patens</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	31
129	PA#1al(200)	115	<i>P. patens</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	15	600	24
130	PA#1am(200)	116	<i>P. fluorescens</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	32
131	PA#1an(200)	117	<i>P. fluorescens</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	35
132	PA#1ao(200)	118	<i>P. fluorescens</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	18	600	41

- 111 -

133	PA#1ap(200)	119	<i>T. thermophila</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	37
134	PA#1aq(200)	120	<i>T. thermophila</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	34
135	PA#1ar(200)	121	<i>T. thermophila</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	600	22
136	PA#1as(200)	122	<i>T. thermophila</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	600	35
137	PAS#1k(200)	123	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	14
138	PAS#1l(200)	124	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	15	600	17
139	PAS#1m(200)	125	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	16
140	PAS#1n(100)	126	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	300	15
141	PAS#1o(200)	127	<i>P. pastoris</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	17
142	PAS#1p(200)	128	<i>P. pastoris</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	17	600	29
143	PAS#1q(200)	129	<i>P. Fluorescens</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	17	600	25
144	PAS#1r(200)	130	<i>P. Fluorescens</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	14
145	PAS#1s(200)	131	<i>P. Fluorescens</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	17	600	24

- 112 -

146	PAS#1t(200)	132	<i>C. glutamicum</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	15
147	PAS#1u(200)	133	<i>C. glutamicum</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	12
148	PAS#1v(200)	134	<i>C. glutamicum</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	11
149	PAS#1w(200)	135	<i>P. patens</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	15
150	PAS#1x(200)	136	<i>P. patens</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	12	600	12
151	PAS#1y(200)	137	<i>P. patens</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	11	600	10
152	PAS#1z(200)	138	<i>K. lactis</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	15
153	PAS#1aa(200)	139	<i>K. lactis</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	15	600	17
154	PAS#1ab(200)	140	<i>K. lactis</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	16
155	PAS#1ac(200)	141	<i>S. cerevisiae</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	14
156	PAS#1ad(200)	142	<i>S. cerevisiae</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	14
157	PAS#1ae(200)	143	<i>S. cerevisiae</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	14
158	PAS#1af(200)	144	<i>T. thermophila</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	17	600	25

- 113 -

159	PAS#1ag(200)	145	<i>T. thermophila</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	17	600	25
160	PAS#1ah(200)	146	<i>T. thermophila</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	15	600	20
161	PAS#1ai(200)	147	<i>H. sapiens</i> (HEK cells)	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	13
162	PAS#1aj(200)	148	<i>H. sapiens</i> (HEK cells)	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	12	600	10
163	PAS#1ak(200)	149	<i>H. sapiens</i> (HEK cells)	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	11
164	PAS#1al(200)	150	<i>B. subtilis</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	12	600	11
165	PAS#1am(200)	151	<i>B. subtilis</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	13
166	PAS#1an(200)	152	<i>B. subtilis</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	14	600	14
167	PA#1at(200)	192	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	31	600	190
168	PA#1au(200)	193	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	26	600	105
169	PAS#1ao(200)	194	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	32	600	211
170	PAS#1ap(200)	195	<i>E. coli</i>	ASPAAPAPASPA APAPSAPA (SEQ ID NO: 1)	26	600	105
B: Assembled low-repetitive nucleotide sequences							
21	PAS#1f/1c/1b(600)	38	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA	14	1800	9

- 114 -

				(SEQ ID NO: 1)			
22	PAS#1d/1f/1c/1b(800)	39	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	2400	8
23	PAS#1h/1e/1i(600)	40	<i>CHO (C. griseus)</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	1800	14
24	PAS#1j/1h/1e/1i(800)	41	<i>CHO (C. griseus)</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	2400	13
25	PA#1d/1c/1b(600)	42	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	1800	15
26	PA#1i/1h/1g/1f(800)	43	<i>CHO (C. griseus)</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	2400	22
27	PA#1e/1d/1c/1b(800)	44	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	2400	14
28	PA#1i/1h/1g/1f/1e/1d/1c/1b(1600)	45	<i>E. coli / CHO (C. griseus)</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	27	4800	24
171	PA#1ae/1c(300)	153	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	900	18
172	PA#1ae/1d(300)	154	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	900	17
173	PA#1d/1c(400)	155	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	1200	17
174	PA#1b/1c/1d(600)	156	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	14	1800	15
175	PA#1d/1b/1c(600)	157	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	20	1800	17
176	PA#1c/1b/1d(600)	158	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	17	1800	16

- 115 -

177	PA#1c/1d/1b(600)	159	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	20	1800	17
178	PA#1b/1d/1c(600)	160	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	1800	16
179	PA#1aa/1e/1d/1c/1b(1000)	161	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	20	3000	17
180	PA#1ab/1aa/1e/1d/1c/1b(1200)	162	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	20	3600	17
181	PA#1ac/1ab/1aa/1e/1d/1c/1b(1400)	163	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	20	4200	16
182	PA#1ad/1ac/1ab/1aa/1e/1d/1c/1b(1600)	164	<i>E. coli</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	20	4800	16
183	PA#1ao/1an/1am(600)	165	<i>P. fluorescens</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	19	1800	27
184	PA#1ai/1ah/1ag/1af(800)	166	<i>C. glutamicum</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	2400	17
185	PA#1y/1x/1w(600)	167	<i>B. subtilis</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	1800	24
186	PA#1j/1k/1l/1m(800)	168	<i>P. pastoris</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	2400	23
187	PA#1p/1o/1n(600)	169	<i>S. cerevisiae</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	18	1800	21
188	PA#1s/1r/1q(600)	170	<i>K. lactis</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	17	1800	23
189	PA#1as/1ar/1aq/1ap(800)	171	<i>T. thermophila</i>	AAPAAPAPAAP AAPAPAAP (SEQ ID NO: 2)	20	2400	30

- 116 -

190	PA#1v/1u/1t(600)	172	<i>H. sapiens</i> (HEK cells)	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	19	1800	28
191	PA#1al/1ak/1j(600)	173	<i>P. patens</i>	AAPAAPAPAAP AAPAPAAPA (SEQ ID NO: 2)	18	1800	24
192	PAS#1n/1b(300)	174	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	900	12
193	PAS#1n/1c(300)	175	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	900	13
194	PAS#1b/1f/1c(600)	176	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	1800	9
195	PAS#1b/1c/1f(600)	177	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	1800	9
196	PAS#1c/1b/1f(600)	178	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	1800	9
197	PAS#1f/1b/1c(600)	179	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	1800	9
198	PAS#1c/1f/1b(600)	180	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	1800	9
199	PAS#1k/1d/1f/1c/1b(1000)	181	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	20	3000	11
200	PAS#1l/1k/1d/1f/1c/1b(1200)	182	<i>E. coli</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	20	3600	12
201	PAS#1s/1q/1r(600)	183	<i>P. fluorescens</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	20	1800	21
202	PAS#1v/1t/1u(600)	184	<i>C. glutamicum</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	17	1800	13

- 117 -

203	PAS#1an/am/1l(600)	185	<i>B. subtilis</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	1800	11
204	PAS#1p/lo/1g(600)	186	<i>P. pastoris</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	17	1800	20
205	PAS#1ae/1ad/1ac(600)	187	<i>S. cerevisiae</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	15	1800	12
206	PAS#1ab/1aa/1z(600)	188	<i>K. lactis</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	17	1800	15
207	PAS#1ah/1ag/1af(600)	189	<i>T. thermophila</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	17	1800	19
208	PAS#1ak/aj/ah(600)	190	<i>H. sapiens</i> (HEK cells)	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	14	1800	10
209	PAS#1y/1x/1w(600)	191	<i>P. patens</i>	ASPAAPAPASP AAPAPSAPA (SEQ ID NO: 1)	17	1800	14

Table 2: Characteristics of prior art nucleotide sequences

	Sequence name	Organism	SEQ ID:	GenBank entry / Patent no.	n _{max}	N _{tot}	NRS
1	PAS#1a(200)	synthetic	11	WO 2008155134	540	600	1 127 680
2	PA#1a(200)	synthetic	14	WO2011144756	540	600	1 127 680
3	PA#3a(200)	synthetic	15	WO2011144756	540	600	1 127 680
4	[(AP) ₅] ₂₀ APA	synthetic	16	US2006/0252120	579	609	1 315 159
5	[AAPAPAPAP] ₁₀ AS module of pBI-SS-(Tom)(AP)51-EGFP	synthetic	17	DQ399411.1	243	276	150 961
6	Large tegument protein	Macacine Herpes virus 1	18	NP_851896.1	197	225	81 858

Example 14: NRS-Calculator, an algorithm to unambiguously identify nucleotide sequence repeats and to calculate the Nucleotide Repeat Score

Generally available software programs such as the Visual Gene Developer (Jung (2011) loc. cit) or the Repfind tool (Betley (2002) loc. cit) do not always work reliably and may require manual corrections in order to calculate all sequence repeats within an analyzed nucleotide sequence properly. In addition, repeats have to be counted manually and the NRS must be calculated separately according to the formula described in Example 13. To provide an algorithm that yields unambiguous results and to facilitate the calculation of the NRS, a simple Python script termed NRS-Calculator is described here. This script, executed on the runtime environment Python 2.7.10 (<http://www.python.org>), is based on a dot matrix sequence comparison and identifies all forward repeats within a potentially long nucleotide sequence, including overlapping repeats, without considering gaps. The dot matrix sequence comparison is a method well known by a person skilled in the art and is described in common bioinformatics text books such as, e.g., Mount (2004) Bioinformatics: Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, 2nd edition, New York.

NRS-Calculator counts the frequencies for each repeat length and automatically calculates the NRS according to the formula described in Example 13. To execute the NRS-Calculator script the runtime environment Python version 2.7.10 was downloaded from <https://www.python.org/downloads> and installed on a ThinkPad L530 notebook (Lenovo, Stuttgart, Germany) running a Windows 7 operating system. The NRS-Calculator script listed below was saved as plain text file designated NRScalculator.py using Microsoft Windows Editor Version 6.1. The nucleotide sequence to be analyzed was saved as FASTA file named sequence.fas within the same folder. Subsequently, the command line shell was opened and the directory containing both the NRScalculator.py and the sequence.fas file was selected. To start the calculation, the following command line was executed:

```
c:\user\admin\NRSfolder> c:\Python27\python.exe NRScalculator.py sequence.fas
```

This command resulted in a screen output of two rows: the left row indicating the repeat length (Length) and the right (second) row indicating the respective repeat frequency (Frequency). In addition, N_{tot} and NRS (number rounded as integer) were stated at the beginning and the end of the output, respectively.

- 119 -

NRS-Calculator script:

```

import math
import sys
class NRSCalculator:
    def __init__(self):
        self.repeats = dict()
        self.sums = dict()
        self.seq = None
        self.range_min = None
        self.range_max = None
    def _match_at(self, row, column):
        return self.seq[row] == self.seq[column]
    def _get_repeats_at(self, row, column):
        length = 1
        search_row = row
        search_column = column
        while True:
            if not 0 <= search_row < len(self.seq):
                break
            if not 0 <= search_column < search_row:
                break
            if length > self.range_max:
                break
            if not self._match_at(search_row, search_column):
                break
            if length >= self.range_min:
                repeats =
self.repeats.setdefault(self.seq[row:row + length], set())
                repeats.add(row)
                repeats.add(column)
                search_row += 1
                search_column += 1
                length += 1
    def _get_repeats(self):
        self.repeats = dict()
        for row in xrange(len(self.seq)):
            for column in xrange(row):
                self._get_repeats_at(row, column)
    def _get_sums(self):
        self.sums = dict()
        for (seq, repeats) in self.repeats.iteritems():
            length = len(seq)
            self.sums[length] = self.sums.get(length, 0) +
len(repeats)
    def set_range(self, range_min, range_max):
        self.range_min = range_min
        self.range_max = range_max
    def set_sequence(self, seq):
        self.seq = seq
    def work(self):
        if not self.seq and not self.range_min and not
self.range_max:

```

- 120 -

```

        raise RuntimeError('Can not work without
initialization')
        self._get_repeats()
        self._get_sums()
        def print_repeats(self):
            print('Sequence (Length bp) : NumRepeats
(Positions)')
            for seq, repeats in sorted(self.repeats.iteritems(),
key=lambda t: len(t[0])):
                list = [seq, len(seq), len(repeats)]
                list.extend(map(lambda value: value + 1,
sorted(repeats)))
                print('%s Ntot = %u : %u (%s)' % (seq, len(seq),
len(repeats), ', '.join(map(lambda value: str(value + 1),
sorted(repeats)))))
            def print_sums(self):
                print('Length\tFrequency')
                for item in self.sums.iteritems():
                    print('%u\t%u' % item)
            def print_score(self):
                sum = 0
                for length, count in self.sums.iteritems():
                    sum += (length ** 2) * math.sqrt(count)
                print('NRS = %.0f' % (sum / len(self.seq)))
def handle_sequence(finder, name, sequence):
    finder.set_range(4, len(sequence))
    finder.set_sequence(sequence)
    finder.work()
    print('%s: Ntot = %u' % (name, len(sequence)))
    #finder.print_repeats()
    finder.print_sums()
    finder.print_score()
if len(sys.argv) != 2:
    print('Usage: %s FILENAME' % sys.argv[0])
    sys.exit(1)
finder = NRSCalculator()
with open(sys.argv[1], 'r') as infile:
    name = 'Unnamed'
    seq = ''
    for line in infile:
        line = line.strip()
        if line.startswith('>'):
            if len(seq) > 0:
                handle_sequence(finder, name, seq)
            name = line
            seq = ''
            continue
        seq += line.upper()
    handle_sequence(finder, name, seq)

```

- 121 -

Exemplary output from NRS-Calculator:

```

>PAS#1b(200): Ntot = 600
Length  Frequency
4       587
5       547
6       478
7       388
8       281
9       158
10      90
11      45
12      6
13      4
14      2
NRS =    13

```

Example 15: Construction of pASK75-PA#1d/1c/1b(600)-IL1Ra, a genetically stable expression vector for the bacterial production of a therapeutic PA#1(600)-IL1Ra fusion protein

For the construction of an expression plasmid encoding the interleukin-1 receptor antagonist (IL-1Ra) as fusion with a 600 residue PA#1 amino acid repeat sequence, the vector pASK75-IL1Ra (**Fig. 4 A**) (SEQ ID NO: 49) was cut with *SapI*, dephosphorylated with shrimp alkaline phosphatase (Thermo Fisher Scientific, Waltham, MA) and ligated with a DNA fragment corresponding to the low repetitive nucleotide sequence cassette encoding the 600 residue PA#1 polypeptide, which was excised from the plasmid pXL1-PA#1d/1c/1b(600) (SEQ ID NO: 79) by restriction digest with *SapI*. After transformation of *E. coli* XL1-Blue (Bullock (1987) loc. cit.), plasmid DNA was prepared and the presence of the inserted DNA fragment was confirmed by restriction analysis and DNA sequencing. The resulting plasmid was designated pASK75-PA#1d/1c/1b(600)-IL1Ra (SEQ ID NO: 77) and is shown in **Fig. 10**.

Example 16: Long-term genetic stability testing of the plasmid pASK75-PA#1b/1c/1b(600)-IL1Ra harboring the low repetitive nucleic acid molecule PA#1d/1c/1b(600) encoding a proline/alanine-rich amino acid repeat sequence

The genetic stability of the plasmid pASK75-PA#1d/1c/1b(600)-IL1Ra (SEQ ID NO: 77) was compared to the genetic stability of pASK75-PA#1a(600)-IL1Ra (SEQ ID NO: 78), a

derivative wherein the PA#1d/1c/1b(600) DNA cassette was replaced by the repetitive nucleic acid PA#1a(600) (SEQ ID NO: 80). To this end, *E. coli* JM83 (Yanisch-Perron C. (1985) loc. cit.) was transformed with the respective plasmid using the calcium chloride method (Sambrook (2001) loc. cit.) and cultured for 7 days at 37 °C, 170 rpm, in 50 ml Luria Bertani (LB) medium supplemented with 100 mg/L ampicillin in a 100 mL shake flask without induction of gene expression. During this period, bacterial cells were twice daily (in the morning and in the evening) transferred into fresh medium using a 1:1000 dilution. On day 7, after a continuous growth over approximately 70 generations, the culture was finally grown to stationary phase and cells were plated on LB/Amp agar. Then, ten individual colonies for each of the two plasmids were picked, each used for inoculation of a 50 mL culture in LB/Amp medium and, after growth to stationary phase over night, plasmid DNA was prepared using the Qiagen Miniprep Kit (Qiagen, Hilden, Germany) and analyzed via *Xba*I/*Hind*III restriction digest (Fig. 11).

Only 6 out of 10 analyzed clones of pASK75-PA#1a(600)-IL1Ra showed the expected bands corresponding to 3093 bp and 2377 bp (Fig. 11 A, lanes 1, 3, 4, 5, 7 and 8). Four clones showed significantly shortened DNA fragments (Fig. 11 A, lanes 2, 6, 9 and 10), indicating deletions within the repetitive PA#1a(600) sequence cassette and, hence, genetic instability. In contrast, all ten analyzed clones of pASK75-PA#1d/1c/1b(600)-IL1Ra revealed the expected bands at 3093 bp and 2377 bp (Fig. 11 B, lanes 1–10), indicating an intact gene cassette encoding proline/alanine-rich amino acid repeat sequences and, thus, high genetic plasmid stability of the low repetitive nucleic acid molecule according to this invention.

Example 17: Construction of genetically stable expression vectors for the bacterial production of human Leptin fused with proline/alanine-rich amino acid repeat sequences.

For the construction of an expression plasmid encoding human Leptin (huLeptin) N-terminally fused with a 600 residue PA#1 amino acid repeat sequence (SEQ ID NO: 82), the vector pASK37-MP-huLeptin (Fig. 12 A) (SEQ ID NO: 81) was cut with *Sap*I, which led to the excision of a small (24 bp) DNA insert containing both *Sap*I recognition sites and a cleaved vector backbone with compatible 5'-GCC/5'-GGC sticky ends at the position directly upstream of the encoded mature N-terminus of human Leptin. These sticky ends are ideally suited for insertion of the low repetitive nucleotide sequence encoding the proline/alanine-rich amino acid repeat sequence at the position directly downstream of the N-terminal start methionine codon (ATG) followed by the proline codon CCA, which was found to allow efficient translational initiation. After isolation of the vector fragment using the QIAquick gel extraction kit (Qiagen, Hilden, Germany) and dephosphorylation with the thermosensitive alkaline phosphatase FastAP (Thermo Fisher Scientific, Waltham, MA), both according to the manufacturers' instructions, the plasmid was ligated with a DNA fragment corresponding to the

low repetitive nucleotide sequence cassette encoding the 600 residue PA#1 polypeptide, which was excised from the plasmid pXL1-PA#1d/1c/1b(600) (SEQ ID NO: 79) by restriction digest with *SapI*. After transformation of *E. coli* XL1-Blue (Bullock (1987) loc. cit.), plasmid DNA was prepared and the presence of the inserted DNA fragment was confirmed by restriction analysis and DNA sequencing. The resulting plasmid was designated pASK37-MP-PA#1d/1c/1b(600)-huLeptin (SEQ ID NO: 82) and is depicted in **Fig. 12 B**. In the same manner, pASK37-MP-PAS#1f/1c/1b(600)-huLeptin (SEQ ID NO: 83), an expression plasmid encoding human Leptin (huLeptin) N-terminally fused with a 600 residue PAS#1 amino acid repeat sequence shown in **Fig. 12 C**, was constructed by insertion of the low repetitive nucleotide sequence PAS#1f/1c/1b (SEQ ID No: 38) excised from pXL1-PAS#1f/1c/1b (SEQ ID No: 84) into the plasmid pASK37-MP-huLeptin (**Fig. 12 A**) (SEQ ID NO: 81). A similar cloning strategy can be applied to construct Leptin versions with C-terminally fused proline/alanine-rich amino acid repeat sequences.

Example 18: Bacterial production, purification and characterization of a fusion protein between a proline/alanine-rich amino acid repeat sequence and a human Leptin mutant encoded on the genetically stable plasmid pASK37-PA#1d/1c/1b(600)hu-Leptin(W100Q)

PA#1(600)-huLeptin(W100Q) a fusion protein between a human Leptin mutant with a tryptophan to glutamine substitution at position 100 of the mature amino acid sequence (UniProtKB accession code P41159) and the proline/alanine-rich amino acid repeat sequence PA#1(600) (SEQ ID NO: 85) (calculated mass: 64.25 kDa) was produced at 30 °C in the cytoplasm of Origami B (Novagene/Merck Millipore, Billerica, MA), an *E. coli* strain which has an oxidizing cytoplasm due to *trxB*, *gor* and *ahpC* mutations (Bessette (1999) Proc. Natl. Acad. Sci. USA 96:13703-13708). To this end, 4 ml LB medium in a sterile 13 mL polypropylene tube (Sarstedt, Nümbrecht, Germany), supplemented with 1 % w/v D-glucose and 100 mg/L ampicillin, was inoculated with a colony of *E. coli* Origami B transformed with the genetically stable expression plasmid pASK37-MP-PA#1d/1c/1b(600)-huLep(W100Q) (SEQ ID NO: 86). Bacterial cells were grown overnight at 30 °C in a shaker at 170 rpm.

Bacterial protein production was performed at 30 °C in a 5 L baffled flask with 2 L terrific broth (TB) medium (Sambrook (2001) loc. cit.) supplemented with 2.5 g/L D-glucose and 100 mg/L ampicillin, which was inoculated with 2 ml of the *E. coli* overnight culture. Bacterial cells were grown at 30°C and recombinant gene expression was induced at OD₅₅₀ = 0.85 by addition of isopropyl-β-D-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM. Bacteria were harvested 19 h after induction, resuspended in 3 ml PBS/E (PBS supplemented with 10 mM EDTA) per 1 g bacterial cell wet weight and lysed using a Panda cell homogenizer (GEA, Parma, Italy). After centrifugation (20,000 rpm, 30 min, 4°C) of the lysate, no inclusion bodies were observed. 1 mM 2,2'-dithiodipyridine was added to the supernatant to boost disulfide

bridge formation in the recombinant Leptin. The supernatant containing the soluble Leptin fusion protein was dialyzed over night at 4°C against a 100-fold volume of PBS. Subsequently, the fusion protein was precipitated at room temperature by dropwise addition of 4 M $(\text{NH}_4)_2\text{SO}_4$ (dissolved in water) under continuous stirring until a final concentration of 1 M $(\text{NH}_4)_2\text{SO}_4$ was reached. After centrifugation for 20 min at 17,000 rpm at room temperature the sediment containing the precipitated PA#1(600)-hu-Leptin(W100/Q) fusion protein was dissolved in PBS and the solution was centrifuged (13,000 rpm, 10 min, room temperature) to remove insoluble contaminants.

The PA#1(600)-hu-Leptin(W100Q) fusion protein was dialyzed twice against 5 L 20 mM Tris/HCl pH 8.5 at 4 °C, each for at least 6 h. Then, the protein solution was subjected to anion exchange chromatography using a 6 ml ResourceQ column (GE Healthcare, Freiburg, Germany) connected to an Äkta purifier system (GE Healthcare, Freiburg, Germany), using 20 mM Tris/HCl pH 8.5 as running buffer. The fusion protein was subsequently eluted using a NaCl concentration gradient. Eluted fractions were collected and further purified via size exclusion chromatography using a Superdex 200 HR10/300 column (GE Healthcare, Freiburg, Germany) equilibrated with PBS.

By this procedure a homogeneous protein preparation without signs of aggregation was obtained with a final yield of 0.8 mg/L bacterial culture. Protein concentration was determined by measuring the absorption at 280 nm using a calculated extinction coefficient (Gill (1989) loc. cit.) of $8605 \text{ M}^{-1} \text{ cm}^{-1}$. SDS-PAGE was performed using a 10 % high molarity Tris buffer system (Fling (1986) loc. cit.) (**Fig. 13 A**). The eluted protein was dialyzed twice against the 10000-fold volume of 10 mM ammonium acetate pH 5.5 and analyzed via ESI mass spectrometry on an maXis Q-TOF instrument (Bruker Daltonics, Bremen, Germany) using the positive ion mode. The deconvoluted spectrum of PA#1(600)-hu-Leptin(W100Q) revealed a mass of 64249.53 Da (**Fig. 13 B**), which essentially coincides with the calculated mass of this fusion protein (64249.80 Da). This clearly demonstrates that a PA#1(600)-hu-Leptin(W100Q) fusion protein, encoded by a genetically stable low repetitive nucleic acid molecule according to this invention, can be produced in *E. coli* in its intact form.

Claims

1. A nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding a polypeptide consisting of at least 100 amino acid residues of proline, alanine and, optionally, serine, wherein said polypeptide forms a random coil, wherein the nucleotide sequence of said nucleic acid has a length of at least 300 nucleotides,
wherein said nucleotide sequence has a Nucleotide Repeat Score (NRS) lower than 1000,
wherein said Nucleotide Repeat Score (NRS) is determined according to the formula:

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{i=1}^{k(n)} f_i(n)}}{N_{tot}},$$

wherein

N_{tot} is the length of said nucleotide sequence,

n is the length of a repeat within said nucleotide sequence, and

$f_i(n)$ is the frequency of said repeat of length n ,

wherein if there is more than one repeat of length n , $k(n)$ is the number of said different sequences of said repeat of length n , otherwise $k(n)$ is 1 for said repeat of length n .

2. The nucleic acid molecule of claim 1, wherein said encoded polypeptide consists of proline and alanine.
3. The nucleic acid molecule of claim 2, wherein said proline residues constitute more than about 10 % and less than about 75 % of said encoded polypeptide.
4. The nucleic acid molecule of claim 1, wherein said encoded polypeptide consists of proline, alanine and serine.
5. The nucleic acid molecule of claim 4, wherein said proline residues constitute more than 4% and less than 40% of said encoded polypeptide.

6. The nucleic acid molecule of any one of claims 1 to 5, wherein said Nucleotide Repeat Score (NRS) is lower than 100.
7. The nucleic acid molecule of any one of claims 1 to 6, wherein said Nucleotide Repeat Score (NRS) is lower than 50.
8. The nucleic acid molecule of any one of claims 1 to 7, wherein said Nucleotide Repeat Score (NRS) is lower than 35.
9. The nucleic acid molecule of any one of claims 1 to 8, wherein the nucleotide sequence of said nucleic acid has a length of at least 900 nucleotides.
10. The nucleic acid molecule of any one of claims 1 to 9, wherein said nucleic acid molecule has an enhanced genetic stability.
11. The nucleic acid molecule of any one of claims 1 to 10, wherein said nucleotide sequence comprises said repeats, wherein said repeats have a maximum length $n_{\max}$, wherein $n_{\max}$ is determined according to the formula:

$$n_{\max} \leq 17 + \frac{N_{\text{tot}}}{600}$$

and wherein N_{tot} is the length of said nucleotide sequence.

12. The nucleic acid molecule of any one of claims 1 to 11, wherein said repeats have a maximum length of about 14, 15, 16, or 17 nucleotides to about 55 nucleotides.
13. The nucleic acid molecule of any one of claims 1 to 12, wherein said encoded polypeptide comprises a repetitive amino acid sequence with a plurality of amino acid repeats, wherein no more than 9 consecutive amino acid residues are identical and wherein said polypeptide forms a random coil.

14. The nucleic acid molecule of any one of claims 1 to 3 and 6 to 13, wherein said nucleic acid molecule is selected from the group consisting of:
- (a) the nucleic acid molecule comprising at least one nucleotide sequence selected from the group consisting of SEQ ID NO: 28, SEQ ID NO: 29, SEQ ID NO: 30, SEQ ID NO: 31, SEQ ID NO: 32, SEQ ID NO: 33, SEQ ID NO: 34, SEQ ID NO: 35, SEQ ID NO: 36, SEQ ID NO: 37, SEQ ID NO: 87, SEQ ID NO: 88, SEQ ID NO: 89, SEQ ID NO: 90, SEQ ID NO: 91, SEQ ID NO: 92, SEQ ID NO: 93, SEQ ID NO: 94, SEQ ID NO: 95, SEQ ID NO: 96, SEQ ID NO: 97, SEQ ID NO: 98, SEQ ID NO: 99, SEQ ID NO: 100, SEQ ID NO: 101, SEQ ID NO: 102, SEQ ID NO: 103, SEQ ID NO: 104, SEQ ID NO: 105, SEQ ID NO: 106, SEQ ID NO: 107, SEQ ID NO: 108, SEQ ID NO: 109, SEQ ID NO: 110, SEQ ID NO: 111, SEQ ID NO: 112, SEQ ID NO: 113, SEQ ID NO: 114, SEQ ID NO: 115, SEQ ID NO: 116, SEQ ID NO: 117, SEQ ID NO: 118, SEQ ID NO: 119, SEQ ID NO: 120, SEQ ID NO: 121, SEQ ID NO: 122, SEQ ID NO: 192 and SEQ ID NO: 193;
 - (b) the nucleic acid molecule comprising the nucleotide sequence consisting of SEQ ID NO: 42, SEQ ID NO: 43, SEQ ID NO: 44, SEQ ID NO: 45, SEQ ID NO: 153, SEQ ID NO: 154, SEQ ID NO: 155, SEQ ID NO: 156, SEQ ID NO: 157, SEQ ID NO: 158, SEQ ID NO: 159, SEQ ID NO: 160, SEQ ID NO: 161, SEQ ID NO: 162, SEQ ID NO: 163, SEQ ID NO: 164, SEQ ID NO: 165, SEQ ID NO: 166, SEQ ID NO: 167, SEQ ID NO: 168, SEQ ID NO: 169, SEQ ID NO: 170, SEQ ID NO: 171, SEQ ID NO: 172, and/or SEQ ID NO: 173;
 - (c) the nucleic acid molecule hybridizing under stringent conditions to the complementary strand of the nucleotide sequence as defined in (a) or (b);
 - (d) the nucleic acid molecule comprising the nucleotide sequence having at least 66.7% identity to the nucleotide sequence as defined in any one of (a), (b) and (c); and
 - (e) the nucleic acid molecule being degenerate as a result of the genetic code to the nucleotide sequence as defined in (a) or (b).

15. The nucleic acid molecule of any one of claims 1 and 4 to 13, wherein said nucleic acid molecule is selected from the group consisting of:
 - (a) the nucleic acid molecule comprising at least one nucleotide sequence selected from the group consisting of SEQ ID NO: 19, SEQ ID NO: 20, SEQ ID NO: 21, SEQ ID NO: 22, SEQ ID NO: 23, SEQ ID NO: 24, SEQ ID NO: 25, SEQ ID NO: 26, SEQ ID NO: 27, SEQ ID NO: 123, SEQ ID NO: 124, SEQ ID NO: 125, SEQ ID NO: 126, SEQ ID NO: 127, SEQ ID NO: 128, SEQ ID NO: 129, SEQ ID NO: 130, SEQ ID NO: 131, SEQ ID NO: 132, SEQ ID NO: 133, SEQ ID NO: 134, SEQ ID NO: 135, SEQ ID NO: 136, SEQ ID NO: 137, SEQ ID NO: 138, SEQ ID NO: 139, SEQ ID NO: 140, SEQ ID NO: 141, SEQ ID NO: 142, SEQ ID NO: 143, SEQ ID NO: 144, SEQ ID NO: 145, SEQ ID NO: 146, SEQ ID NO: 147, SEQ ID NO: 148, SEQ ID NO: 149, SEQ ID NO: 150, SEQ ID NO: 151, SEQ ID NO: 152, SEQ ID NO: 194 and SEQ ID NO: 195;
 - (b) the nucleic acid molecule comprising the nucleotide sequence selected from the group consisting of SEQ ID NO: 38, SEQ ID NO: 39, SEQ ID NO: 40, SEQ ID NO: 41, SEQ ID NO: 174, SEQ ID NO: 175, SEQ ID NO: 176, SEQ ID NO: 177, SEQ ID NO: 178, SEQ ID NO: 179, SEQ ID NO: 180, SEQ ID NO: 181, SEQ ID NO: 182, SEQ ID NO: 184, SEQ ID NO: 185, SEQ ID NO: 186, SEQ ID NO: 187, SEQ ID NO: 188, SEQ ID NO: 189, SEQ ID NO: 190, and SEQ ID NO: 191;
 - (c) the nucleic acid molecule hybridizing under stringent conditions to the complementary strand of the nucleotide sequence as defined in (a) or (b);
 - (d) the nucleic acid molecule comprising the nucleotide sequence having at least 56% identity to the nucleotide sequence as defined in any one of (a), (b) and (c);
 - (e) the nucleic acid molecule being degenerate as a result of the genetic code to the nucleotide sequence as defined in (a) or (b).
16. The nucleic acid molecule of any one of claims 1 to 15 operably linked in the same reading frame to a nucleic acid encoding a biologically active protein.
17. The nucleic acid molecule of claim 16, wherein said biologically active protein is a therapeutically effective protein.

18. The nucleic acid molecule of claim 16 or 17, wherein said biologically active protein is selected from the group consisting of a binding protein, an antibody fragment, a cytokine, a growth factor, a hormone, an enzyme, a protein vaccine, a peptide vaccine, a peptide which consists of up to 50 amino acid residues or a peptidomimetic.
19. The nucleic acid molecule of claim 18, wherein said binding protein is selected from the group consisting of antibodies, Fab fragments, Fab' fragments, F(ab')₂ fragments, single chain variable fragments (scFv), (single) domain antibodies, isolated variable regions of antibodies (VL and/or VH regions), CDRs, immunoglobulin domains, CDR-derived peptidomimetics, lectins, protein scaffolds, fibronectin domains, tenascin domains, protein A domains, SH3 domains, ankyrin repeat domains, and lipocalins.
20. The nucleic acid molecule of any one of claims 16 to 18, wherein said biologically active protein is selected from the group consisting of interleukin 1 receptor antagonist, leptin, acid sphingomyelinase, adenosine deaminase, agalsidase alfa, alpha-1 antitrypsin, alpha atrial natriuretic peptide, alpha-galactosidase, alpha-glucosidase, alpha-N-acetylglucosaminidase, alteplase, amediase, amylin, amylin analog, anti-HIV peptide fusion inhibitor, arginine deiminase, asparaginase, B domain deleted factor VIII, bone morphogenetic protein, bradykinin antagonist, B-type natriuretic peptide, bouganin, growth hormone, chorionic gonadotropin, CD3 receptor antagonist, CD19 antagonist, CD20 antagonist, CD40 antagonist, CD40L antagonist, cerebroside sulfatase, coagulation factor VIIa, coagulation factor XIII, coagulation factor IX, coagulation factor X, complement component C3 inhibitor, complement component 5a antagonist, C-peptide, CTLA-4 antagonist, C-type natriuretic peptide, defensin, deoxyribonuclease I, EGFR receptor antagonist, epidermal growth factor, erythropoietin, exendin-4, ezrin peptide 1, FcγIIB receptor antagonist, fibroblast growth factor 21, follicle-stimulating hormone, gastric inhibitory polypeptide (GIP), GIP analog, glucagon, glucagon receptor agonist, glucagon-like peptide 1 (GLP-1), GLP-1 analog, glucagon-like peptide 2 (GLP-2), GLP-2 analog, gonadorelin, gonadotropin-releasing hormone agonist, gonadotropin-releasing hormone antagonist, gp120, gp160, granulocyte colony stimulating factor (G-CSF), granulocyte macrophage colony

stimulating factor (GM-CSF), grehlin, grehlin analog, growth hormone, growth hormone-releasing hormone, hematide, hepatocyte growth factor, hepatocyte growth factor receptor (HGFR) antagonist, hepcidin antagonist, hepcidin mimetic, Her2/neu receptor antagonist, histrelin, hirudin, hsp70 antagonist, humanin, hyaluronidase, hydrolytic lysosomal glucocerebroside-specific enzyme, iduronate-2-sulfatase, IgE antagonists, insulin, insulin analog, insulin-like growth factor 1, insulin-like growth factor 2, interferon-alpha, interferon-alpha antagonist, interferon-alpha superagonist, interferon-alpha-n3, interferon-beta, interferon-gamma, interferon-lambda, interferon tau, interleukin, interleukin 2 fusion protein, interleukin-22 receptor subunit alpha (IL-22ra) antagonist, irisin, islet neogenesis associated protein, keratinocyte growth factor, Kv1.3 ion channel antagonists, lanthipeptide, lipase, luteinizing hormone, lutropin alpha, lysostaphin, mannosidase, N-acetylgalactosamine-6-sulfatase, N-acetylglucosaminidase, neutrophil gelatinase-associated lipocalin, octreotide, ω -conotoxin, Ornithodoros moubata complement inhibitor, osteogenic protein-1, osteoprotegerin, oxalate decarboxylase, P128, parathyroid hormone, Phylomer, PD-1 antagonist, PDGF antagonist, phenylalanine ammonia lyase, platelet derived growth factor, proinsulin, protein C, relaxin, relaxin analog, secretin, RGD peptide, ribonuclease, senrebotase, serine protease inhibitor, soluble complement receptor type 1, soluble DCC receptor, soluble TACI receptor, soluble tumor necrosis factor I receptor (sTNF-RI), soluble tumor necrosis factor II receptor (sTNF-RII), soluble VEGF receptor Flt-1, soluble Fc γ IIB receptor, somatostatin, somatostatin analog, streptokinase, T-cell receptor ligand, tenecteplase, teriparatide, thrombomodulin alpha, thymosin alpha 1, toll like receptor inhibitor, tumor necrosis factor (TNF α), tumor necrosis factor α antagonist, uricase, vasoactive intestinal peptide, vasopressin, vasopressin analog, VEGF antagonist, von Willebrand factor.

21. A vector comprising the nucleic acid molecule of any one of claims 1 to 20.
22. A host or host cell comprising the nucleic acid molecule of any one of claims 1 to 20, a host or host cell comprising the vector of claim 21, or a host or host cell transformed with a vector of claim 21.

23. A method for preparing the nucleic acid molecule of any one of claims 1 to 20, wherein the method comprises culturing the host or host cell of claim 22 and optionally isolating the produced nucleic acid molecule.
24. A method for preparing the vector of claim 21, wherein the method comprises culturing the host or host cell of claim 22 and optionally isolating the produced vector.
25. A method for preparing a polypeptide encoded by the nucleic acid molecule of any one of claims 1 to 20, wherein the method comprises culturing/raising the host or host cell of claim 22 and optionally isolating the produced polypeptide.
26. A method for preparing a drug conjugate, wherein said drug conjugate comprises the polypeptide encoded by the nucleic acid molecule of any one of claims 1 to 15 and further comprises (i) a biologically active protein and/or (ii) a small molecule and/or (iii) a carbohydrate, wherein the method further comprises culturing the host or host cell of claim 22 and optionally isolating the produced polypeptide and/or drug conjugate.
27. The method for preparing the drug conjugate of claim 26, wherein said biologically active protein is a therapeutically effective protein.
28. The method for preparing the drug conjugate of claim 26 or 27, wherein said biologically active protein is selected from the group consisting of a binding protein, an antibody fragment, a cytokine, a growth factor, a hormone, an enzyme, a protein vaccine, a peptide vaccine, a peptide which consists of up to 50 amino acid residues or a peptidomimetic.
29. The method for preparing the drug conjugate of claim 28, wherein said binding protein is selected from the group consisting of antibodies, Fab fragments, Fab' fragments, F(ab')₂ fragments, single chain variable fragments (scFv), (single) domain antibodies, isolated variable regions of antibodies (VL and/or VH regions), CDRs, immunoglobulin domains, CDR-derived peptidomimetics, lectins, protein scaffolds, fibronectin domains, tenascin domains, protein A domains, SH3 domains, ankyrin repeat domains, and lipocalins.

30. The method for preparing the drug conjugate of any one of claims 26 to 28, wherein said biologically active protein is selected from the group consisting of interleukin 1 receptor antagonist, leptin, acid sphingomyelinase, adenosine deaminase, agalsidase alfa, alpha-1 antitrypsin, alpha atrial natriuretic peptide, alpha-galactosidase, alpha-glucosidase, alpha-N-acetylglucosaminidase, alteplase, amediase, amylin, amylin analog, anti-HIV peptide fusion inhibitor, arginine deiminase, asparaginase, B domain deleted factor VIII, bone morphogenetic protein, bradykinin antagonist, B-type natriuretic peptide, bouganin, growth hormone, chorionic gonadotropin, CD3 receptor antagonist, CD19 antagonist, CD20 antagonist, CD40 antagonist, CD40L antagonist, cerebroside sulfatase, coagulation factor VIIa, coagulation factor XIII, coagulation factor IX, coagulation factor X, complement component C3 inhibitor, complement component 5a antagonist, C-peptide, CTLA-4 antagonist, C-type natriuretic peptide, defensin, deoxyribonuclease I, EGFR receptor antagonist, epidermal growth factor, erythropoietin, exendin-4, ezrin peptide 1, FcγIIB receptor antagonist, fibroblast growth factor 21, follicle-stimulating hormone, gastric inhibitory polypeptide (GIP), GIP analog, glucagon, glucagon receptor agonist, glucagon-like peptide 1 (GLP-1), GLP-1 analog, glucagon-like peptide 2 (GLP-2), GLP-2 analog, gonadorelin, gonadotropin-releasing hormone agonist, gonadotropin-releasing hormone antagonist, gp120, gp160, granulocyte colony stimulating factor (G-CSF), granulocyte macrophage colony stimulating factor (GM-CSF), grehlin, grehlin analog, growth hormone, growth hormone-releasing hormone, hematide, hepatocyte growth factor, hepatocyte growth factor receptor (HGFR) antagonist, hepcidin antagonist, hepcidin mimetic, Her2/neu receptor antagonist, histrelin, hirudin, hsp70 antagonist, humanin, hyaluronidase, hydrolytic lysosomal glucocerebroside-specific enzyme, iduronate-2-sulfatase, IgE antagonists, insulin, insulin analog, insulin-like growth factor 1, insulin-like growth factor 2, interferon-alpha, interferon-alpha antagonist, interferon-alpha superagonist, interferon-alpha-n3, interferon-beta, interferon-gamma, interferon-lambda, interferon tau, interleukin, interleukin 2 fusion protein, interleukin-22 receptor subunit alpha (IL-22ra) antagonist, irisin, islet neogenesis associated protein, keratinocyte growth factor, Kv1.3 ion channel antagonists, lanthipeptide, lipase, luteinizing hormone, lutropin alpha, lysostaphin, mannosidase, N-acetylgalactosamine-6-sulfatase, N-

acetylglucosaminidase, neutrophil gelatinase-associated lipocalin, octreotide, ω -conotoxin, *Ornithodoros moubata* complement inhibitor, osteogenic protein-1, osteoprotegerin, oxalate decarboxylase, P128, parathyroid hormone, Phylomer, PD-1 antagonist, PDGF antagonist, phenylalanine ammonia lyase, platelet derived growth factor, proinsulin, protein C, relaxin, relaxin analog, secretin, RGD peptide, ribonuclease, senrebotase, serine protease inhibitor, soluble complement receptor type 1, soluble DCC receptor, soluble TACI receptor, soluble tumor necrosis factor I receptor (sTNF-RI), soluble tumor necrosis factor II receptor (sTNF-RII), soluble VEGF receptor Flt-1, soluble Fc γ IIB receptor, somatostatin, somatostatin analog, streptokinase, T-cell receptor ligand, tenecteplase, teriparatide, thrombomodulin alpha, thymosin alpha 1, toll like receptor inhibitor, tumor necrosis factor (TNF α), tumor necrosis factor α antagonist, uricase, vasoactive intestinal peptide, vasopressin, vasopressin analog, VEGF antagonist, von Willebrand factor.

31. The method for preparing the drug conjugate of claims 26 to 30, wherein said small molecule is selected from the group consisting of angiogenesis inhibitors, anti-allergic drugs, anti-emetic drugs, anti-depressant drugs, anti-hypertensive drugs, anti-inflammatory drugs, anti-infective drugs, anti-psychotic drugs, anti-proliferative (cytotoxic and cytostatic) drugs, calcium antagonists and other circulatory organ drugs, cholinergic agonists, drugs acting on the central nervous system, drugs acting on the respiratory system, hormones, steroids, polyketides, carbohydrates, oligosaccharides, nucleic acids, nucleic acid derivatives, antisense nucleic acids, small interference RNAs (siRNAs), micro RNA (miR) inhibitors, microRNA mimetics, DNA aptamers and RNA aptamers.
32. A method for selecting a genetically stable nucleic acid molecule, wherein said nucleic acid molecule comprises a nucleotide sequence encoding a polypeptide consisting of at least 100 amino acid residues of proline, alanine and, optionally, serine, wherein said polypeptide forms a random coil, wherein said nucleotide sequence has a length of at least 300 nucleotides,
the method comprising a step of selecting a nucleic acid molecule comprising a nucleotide sequence having a Nucleotide Repeat Score (NRS) lower than 1000,

wherein said Nucleotide Repeat Score (NRS) is determined according to the formula:

$$NRS = \frac{\sum_{n=4}^{N_{tot}-1} n^2 \sqrt{\sum_{i=1}^{k(n)} f_i(n)}}{N_{tot}},$$

wherein

N_{tot} is the length of said nucleotide sequence,

n is the length of a repeat within said nucleotide sequence, and

$f_i(n)$ is the frequency of said repeat of length n ,

wherein if there is more than one repeat of length n , $k(n)$ is the number of said different sequences of said repeat of length n , otherwise $k(n)$ is 1 for said repeat of length n .