

Titel**VERFAHREN ZUR OPTIMIERUNG EINER NUKLEOTIDSEQUENZ DURCH
AUSTAUSCH SYNONYMER CODONS FÜR DIE EXPRESSION EINER
AMINOSÄURESEQUENZ IN EINEM ZIELORGANISMUS**

5

Verweis auf frühere Anmeldung

Diese Anmeldung beansprucht die Priorität der deutschen
Patentanmeldung Nr. 10 2022 118 459.5, eingereicht am 22. Juli
10 2022, deren Offenbarungsgehalt hiermit durch Verweis
vollumfänglich einbezogen wird.

Sequenzprotokoll

15 Diese Anmeldung beinhaltet ein elektronisches Sequenzprotokoll
im xml-Format nach WIPO ST.26 Standard mit 21 Sequenzen als Teil
der Beschreibung, auf dessen Inhalte hiermit vollumfänglich
Bezug genommen wird.

20 Gebiet der Erfindung

Die vorliegende Erfindung liegt auf dem Gebiet der Erzeugung
synthetischer Nukleotidsequenzen und deren Verwendung zur
Erzeugung von Proteinen, indem diese Nukleotidsequenzen in ein
25 Expressionssystem mit einem geeigneten Wirtsorganismus
eingebracht werden, der das durch die Nukleotidsequenz codierte
Protein exprimiert. Die vorliegende Erfindung betrifft
insbesondere Verfahren, bei denen eine Nukleotidsequenz für die
Expression in einem vorbestimmten Wirtsorganismus optimiert
30 wird.

Hintergrund der Erfindung

Heterologe Expressionssysteme besitzen eine große Bedeutung in der Biotechnologie, um rekombinante Proteine herzustellen. Als „Expressionssystem“ wird ein biologisches System bezeichnet, das in der Lage ist, gezielt und kontrolliert Proteinbiosynthese zu betreiben, d. h. bestimmte Proteine nach der Vorlage einer Nukleotidsequenz herzustellen, also zu „exprimieren“.

Unter „heterologer Expression“ wird hierbei die Expression eines Gens oder eines Teils davon in einem Wirtsorganismus verstanden, der dieses Gen oder Genfragment von Natur aus nicht besitzt. Die entsprechenden Nukleotidsequenzen werden durch Gentechnologie bzw. rekombinante DNA-Technologie, z. B. mithilfe von Vektoren oder Genom-Editierung, in den Wirtsorganismus eingebracht, woraufhin dieser vermehrt und zur Überproduktion des Proteins angeregt wird. Eine „homologe Expression“ bezieht sich demzufolge auf die Expression eines Gens oder eines Teils davon in einem Wirtsorganismus bzw. einem System, aus dem es ursprünglich stammt.

Eine heterologe Proteinexpression kann in vielen Arten von Wirtsorganismen durchgeführt werden. Der Wirtsorganismus kann z. B. ein Bakterium, ein Pilz, eine Hefe, eine Insektenzelle, eine Säugetierzelle oder eine Pflanzenzelle sein. Ein häufig auftretendes Problem bei der heterologen Proteinexpression ist eine geringe Transkriptions- und Translationsrate einer artfremden Nukleotidsequenz in einem bestimmten Wirtsorganismus, hier auch als Zielorganismus bezeichnet. Ursächlich ist u. a. die Degeneration des genetischen Codes, die dazu führt, dass für die meisten der translational einzubauenden Aminosäuren jeweils mehrere Codons mit gleicher Bedeutung, hier auch als „synonyme Codons“ bezeichnet, verfügbar sind. Ein Codon ist eine Abfolge von drei aufeinanderfolgenden Nukleobasen einer Nukleinsäuresequenz, d. h. ein „Basentriplett“, das für eine

Aminosäure codieren kann. Insgesamt existieren 64 mögliche Codons, davon codieren 61 für die insgesamt 20 kanonischen der proteinogenen Aminosäuren, drei weitere codieren für Stop-Codons.

5

Unterschiedliche Organismen können Codons für die Expression einer Aminosäure mit unterschiedlicher Häufigkeit verwenden, auch als „Codon Usage“ oder „Codon Bias“ bezeichnet. Darüber hinaus hat sich gezeigt, dass Codons in den Genen eines

10 Organismus nicht zufällig angeordnet sind, sondern vielmehr die beobachteten Frequenzen von Codonpaaren widererwarten vom Produkt der jeweiligen Einzelfrequenzen abweichen können und dadurch statistisch „unterrepräsentiert“ oder „überrepräsentiert“ sind. Dieser Zusammenhang wird auch als
15 „Codonkontext“ bezeichnet, welcher einen zusätzlichen Einfluss auf die Translationseffizienz ausüben kann.

Bekannte Möglichkeiten zur Steigerung der Expressionsrate sind daher die Codonoptimierung durch Austausch einzelner Codons der

20 zu exprimierenden Nukleotidsequenz gegen synonyme Codons mit höheren Frequenzen im Zielorganismus (auch als „Codon-Optimization“ oder „Codon-Usage-Optimization“ bezeichnet), sowie durch Übertragung der Frequenzen von Codons stark exprimierter Gene im Zielorganismus auf das Zielprotein (auch als „Codon
25 Adaptation Index“ bezeichnet). Hinzu kommt die Möglichkeit der Codonkontext-Optimierung, wobei die Codons der zu exprimierenden Nukleinsäuresequenz durch Austausch mit überrepräsentierten oder unterrepräsentierten synonymen Codons an den Codonkontext des Zielorganismus angepasst werden, ohne die codierte
30 Aminosäuresequenz zu verändern.

WO 2020/024917 A1 offenbart ein computerimplementiertes Verfahren, bei dem Nukleinsäuresequenzen u. a. auf der Basis des

Codon Adaptation Index und des Codonkontextes unter Verwendung eines computergestützten NSGA-III-Algorithmus für die Expression eines Proteins in einem Wirt optimiert werden.

5 Aus WO 2004/059556 A2 ist ein computergestütztes Verfahren bekannt, bei dem eine vorgegebene Nukleinsäuresequenz mithilfe einer Gütefunktion für die Expression in einem vorgegebenen Zielorganismus optimiert wird. Die Gütefunktion kann als Gütekriterium unter anderem die Codonverwendung und den
10 Codonkontext berücksichtigen.

In der WO 2008/000632 A1 wird ein weiteres Verfahren vorgeschlagen, bei dem aus einer vorgegebenen Nukleotidsequenz, die eine vorbestimmte Aminosäuresequenz codiert, durch
15 Austauschen von einem oder mehreren synonymen Codons in mehreren Wiederholungsschritten neue codierende Sequenzen erzeugt und anhand eines Fitnesswertes, der u. a. den Codonkontext des Wirtsorganismus berücksichtigt, weiterentwickelt werden.

20 Aus der WO 2018/104385 A1 ist ein Verfahren zur Ermittlung einer optimierten Nukleotidsequenz, die eine vorbestimmte Aminosäuresequenz codiert und für die Expression in einem bestimmten Zielorganismus optimiert ist, bekannt, wobei eine Vielzahl von Kandidaten-Nukleotidsequenzen erzeugt und unter
25 Verwendung eines statistischen maschinellen Lernalgorithmus bewertet werden.

Ein weiteres auf dem Codonkontext des Zielorganismus basierendes Verfahren zur Optimierung von Nukleotidsequenzen für die
30 heterologe Proteinexpression ist aus WO 2007/130650 A2 bekannt.

Es hat sich gezeigt, dass diese konventionellen Optimierungen oft nicht ausreichen, um eine wesentliche Verbesserung der

Proteinausbeute im heterologen System zu erreichen. Darüber hinaus sind die bekannten Verfahren in der Regel auf eine mengenmäßige Optimierung der transkribierten mRNA oder der Proteinausbeute abgestimmt und lassen dabei die Löslichkeit des exprimierten Proteins unberücksichtigt, die jedoch für dessen Verwertbarkeit für wissenschaftliche, medizinisch-pharmazeutische und industrielle Zwecke wesentlich ist.

Es ist daher eine Aufgabe der vorliegenden Erfindung, verbesserte Verfahren und Vorrichtungen zur Optimierung einer Nukleotidsequenz für die Expression einer vorgegebenen Aminosäuresequenz in einem Zielorganismus bereitzustellen, welche die oben genannten Probleme zumindest teilweise lösen und dabei insbesondere den Anteil an löslichem Protein in der heterologen Proteinexpression steigern können.

Diese Aufgabe wird durch die Gegenstände der unabhängigen Ansprüche gelöst. Bevorzugte Ausführungsformen der Erfindung sind Gegenstände der abhängigen Ansprüche und der nachfolgenden Beschreibung.

Beschreibung der Erfindung

Gegenstand der vorliegenden Erfindung ist ein Verfahren zur Optimierung einer Nukleotidsequenz für die Expression einer vorgegebenen Aminosäuresequenz in mindestens einem Zielorganismus. Die Expression kann prinzipiell heterolog oder homolog sein. Vorzugsweise handelt es sich um eine heterologe Expression.

Die Nukleotidsequenz umfasst eine Vielzahl von Basentripletts, wobei an mindestens einer Änderungsposition der Nukleotidsequenz ein Basentriplett, das eine Aminosäure der vorgegebenen

Aminosäuresequenz codiert, durch ein synonymes Basentriplett, das dieselbe Aminosäure der vorgegebenen Aminosäuresequenz codiert, ersetzt wird, um die Nukleotidsequenz für die Expression in dem mindestens einen Zielorganismus zu optimieren.

5

Eine erfindungsgemäße Änderungsposition umfasst hierbei eine direkte Aufeinanderfolge von n Basentriplets, die ein erstes Codon-n-Tupel bildet und einen Sequenzabschnitt von n Aminosäuren der vorgegebenen Aminosäuresequenz codiert, der ein Aminosäure-n-Tupel bildet, wobei das Aminosäure-n-Tupel mit einer vorbestimmten Menge von Aminosäure-n-Tupel-Ereignissen im Genom des mindestens einen Zielorganismus oder eines Teils davon und/oder in Genomen von zur Infektion des mindestens einen Zielorganismus befähigten Viren oder Teilen davon codiert wird.

15

Das erfindungsgemäße Verfahren beinhaltet, dass mindestens eines der n Basentriplets aus der direkten Aufeinanderfolge der mindestens einen Änderungsposition durch ein synonymes Basentriplett ersetzt wird, wobei das synonyme Basentriplett so gewählt wird, dass ein zweites Codon-n-Tupel resultiert, welches in Bezug auf die Menge der Aminosäure-n-Tupel-Ereignisse eine höhere relative Codon-n-Tupel-Häufigkeit in dem Genom oder des Teils davon des mindestens einen Zielorganismus und/oder in den Genomen oder den Teilen davon der zur Infektion des mindestens einen Zielorganismus befähigten Viren aufweist als das erste Codon-n-Tupel.

20

25

Hierbei ist n eine natürliche Zahl größer gleich zwei und insbesondere kleiner gleich einer Gesamtzahl N der Aminosäuren der vorgegebenen Aminosäuresequenz.

30

Der Erfindung liegt die Erkenntnis zugrunde, dass sich der Einfluss einer direkten Aufeinanderfolge von n Basentriplets,

hier als Codon-n-Tupel bezeichnet, auf die Translationseffizienz einer durch die n Basentriplets codierten Aufeinanderfolge von n Aminosäuren, hier als Aminosäure-n-Tupel bezeichnet, in einem bestimmten Zielorganismus durch die relative Häufigkeit

5 ausdrücken lässt, mit der die direkte Aufeinanderfolge von n Basentriplets die direkte Aufeinanderfolge von n Aminosäuren innerhalb des Genoms eines Wirtsorganismus codiert, hier als relative Codon-n-Tupel Häufigkeit bezeichnet.

10 Ohne Beschränkung auf theoretische Überlegungen wird hierbei angenommen, dass der Translationsprozess zumindest zeitweise von zwei oder mehr aufeinanderfolgenden Basentriplets beeinflusst wird, die während der Translation gleichzeitig am Ribosom gebunden sind. Dabei haben die Erfinder Anhaltspunkte dafür

15 gefunden, dass sich in Genomen häufiger Codon-n-Tupel mit vorteilhaftem Einfluss als synonyme Codon-n-Tupel mit weniger vorteilhaftem oder ungünstigem Einfluss finden lassen.

Mit anderen Worten haben die Erfinder erkannt, dass das

20 Verhältnis aus der absoluten Häufigkeit eines Codon-n-Tupels und der absoluten Häufigkeit des entsprechenden Aminosäure-n-Tupels, welches durch das Codon-n-Tupel codiert wird, ein vorteilhaftes Maß ist, um die Eignung einer vorgegebenen Nukleinsäuresequenz für die Expression in einem bestimmten Zielorganismus zu

25 quantifizieren. Die relative Häufigkeit kann Werte zwischen 0 und 1 bzw. 0% und 100% annehmen, wobei die relative Häufigkeit gleich 0 ist, wenn ein bestimmtes Codon-n-Tupel im Genom oder des Teils davon des Zielorganismus überhaupt nicht zur Codierung des entsprechenden Aminosäure-n-Tupels verwendet wird. Die

30 relative Häufigkeit ist gleich 1 bzw. 100%, wenn ein bestimmtes Aminosäure-n-Tupel im Genom oder des Teils davon des Zielorganismus ausschließlich durch ein bestimmtes Codon-n-Tupel codiert wird. Auf diese Weise ist es möglich, eine Nukleinsäure

durch gezieltes Ersetzen von Codon-n-Tupeln mit geringer relativer Häufigkeit durch Codon-n-Tupel mit höherer relativer Häufigkeit im Genom eines vorbestimmten Zielorganismus zu optimieren, sodass diese dem Zielorganismus für die Expression
5 einer vorgegebenen Aminosäuresequenz besser angepasst ist.

Wenn ein bestimmtes Aminosäure-n-Tupel der Aminosäuresequenz im Genom des Zielorganismus oder des Teils davon bzw. in den Genomen der Viren oder der Teile davon überhaupt nicht vorkommt,
10 kann im Prinzip ein beliebiges Codon-n-Tupel ausgewählt werden, welches dieses Aminosäure-n-Tupel codiert. Beispielsweise kann den zugehörigen synonymen Codon-n-Tupeln jeweils eine einheitliche relative Häufigkeit von $1/i$ zugewiesen werden, wobei i die Anzahl der synonymen Codon-n-Tupel ist, die das
15 Aminosäure-n-Tupel codieren. Alternativ kann Codon-n-Tupeln, welche dieses Aminosäure-n-Tupel codieren, eine relative Häufigkeit von 0 zugewiesen werden. Eine weitere Möglichkeit besteht darin, diese Änderungsposition bzw. dieses Aminosäure-n-Tupel von der Optimierung auszuschließen.

20

Es hat sich gezeigt, dass sich bei einem gegebenen Zielorganismus mithilfe des erfindungsgemäßen Verfahrens die heterologe Proteinexpression und insbesondere die Ausbeute an löslichem Protein um ein Vielfaches gegenüber den bisher
25 bekannten Verfahren steigern lässt. In diesem Zusammenhang wird auch auf die Ergebnisse der Vergleichsversuche in den nachfolgenden Ausführungsbeispielen verwiesen.

Darüber hinaus kann das erfindungsgemäße Verfahren auch die
30 Biosynthese von Proteinen ermöglichen, die sich bisher nicht oder kaum heterolog exprimieren ließen. Auf diese Weise führt das erfindungsgemäße Verfahren zu einer Verbesserung der Effizienz und Nachhaltigkeit in der biotechnologischen

Proteinherstellung für wissenschaftliche, medizinische und technische bzw. industrielle Zwecke.

In bevorzugten Ausführungsformen des Verfahrens ist n kleiner
5 gleich 50, kleiner gleich 40, kleiner gleich 30, kleiner gleich
20 oder kleiner gleich 10. Besonders bevorzugt ist n ausgewählt
aus der Gruppe bestehend aus $n = 2$, $n = 3$, $n = 4$, $n = 5$, $n = 6$
und beliebigen Kombinationen davon. In einer Ausführungsform ist
 $n = 2$. In einer Ausführungsform ist $n = 3$. In einer
10 Ausführungsform ist $n = 4$. in einer Ausführungsform ist $n = 5$.
In einer Ausführungsform ist $n = 6$. Besonders bevorzugt ist n
größer gleich drei. Mit diesen n wird vorteilhafte Anpassung der
Nukleotidsequenz für die Expression in dem Zielorganismus
erreicht.

15 So, wie der Begriff hier verwendet wird, bedeutet „vorbestimmt“,
dass die Menge von Aminosäure- n -Tupel-Ereignissen durch das
Genom bzw. das Proteom des mindestens einen Zielorganismus oder
eines Teils davon bzw. die Genome von zur Infektion des
20 mindestens einen Zielorganismus befähigten Viren oder Teilen
davon vorgegeben ist. Es versteht sich daher, dass die
Bestimmung der Anzahl von Ereignissen, nachfolgend auch als
absolute Häufigkeit bezeichnet, mit der ein bestimmtes
Aminosäure- n -Tupel im Genom des mindestens einen Zielorganismus
25 oder eines Teils bzw. in den Genomen von zur Infektion des
mindestens einen Zielorganismus befähigten Viren oder Teilen
davon codiert wird, einen Schritt darstellt, der während des
Verfahrens durchgeführt werden kann, aber nicht durchgeführt
werden muss. Vielmehr kann die Information über die absolute
30 Häufigkeit, mit der ein Aminosäure- n -Tupel im Genom des
mindestens einen Zielorganismus oder eines Teils davon und/oder
in Genomen von zur Infektion des mindestens einen Zielorganismus
befähigten Viren oder Teilen davon codiert wird, auch auf

anderem Wege, z. B. aus Datenbanken oder dergleichen, in das erfindungsgemäße Verfahren einbezogen werden. Selbstverständlich ist es aber auch möglich, dass die Bestimmung der Menge von Ereignissen, mit denen das Aminosäure-n-Tupel im Genom des

5 mindestens einen Zielorganismus oder eines Teils davon bzw. in Genomen von zur Infektion des mindestens einen Zielorganismus befähigten Viren oder Teilen davon codiert wird, als Verfahrensschritt durchgeführt wird.

10 Gleiches gilt im Grunde genommen auch für die Bestimmung der Anzahl von Ereignissen, d. h. der absoluten Häufigkeit, mit denen ein bestimmtes Codon-n-Tupel im Genom des mindestens einen Zielorganismus oder eines Teils davon bzw. in Genomen von zur Infektion des mindestens einen Zielorganismus befähigten Viren

15 oder Teilen davon vorkommt, und/oder für die sich daraus ergebende erfindungsgemäße relative Häufigkeit des Codon-n-Tupels. Beispielsweise ist es möglich, dass die absolute Häufigkeit und/oder die relative Häufigkeit im Wesentlichen jedes kombinatorisch möglichen Codon-n-Tupels im Genom des

20 mindestens einen Zielorganismus oder eines Teils davon bzw. in Genomen von zur Infektion des mindestens einen Zielorganismus befähigten Viren oder Teilen davon in einer Datenbank hinterlegt sind und in Form einer Datenbankinformation in das erfindungsgemäße Verfahren einbezogen werden.

25 Die Wahl des synonymen Basentriplets erfolgt in dem erfindungsgemäßen Verfahren gezielt unter dem Gesichtspunkt, dass das zweite Codon-n-Tupel die erfindungsgemäß erforderliche Bedingung der höheren relativen Codon-n-Tupel-Häufigkeit

30 gegenüber dem ersten Codon-n-Tupel erfüllt. Das Auswählen des synonymen Basentriplets kann daher insbesondere ein Ermitteln und/oder Bewerten der relativen Codon-n-Tupel-Häufigkeit des zweiten Codon-n-Tupels umfassen. Der Begriff „Ermitteln“ kann,

wie oben ausgeführt, z. B. in Form einer Berechnung der relativen Codon-n-Tupel-Häufigkeit des zweiten Codon-n-Tupels oder auch in Form eines Datenabgleichs, der Einbeziehung einer Datenbankinformation oder dergleichen durchgeführt werden. Eine
5 bevorzugte Verfahrensdurchführung umfasst daher zumindest die folgenden Schritte: a) Bestimmen der mindestens einen Änderungsposition; b) Ersetzen des mindestens einen Basentriplets der mindestens einen Änderungsposition durch das synonyme Basentriplett; c) Ermitteln der relativen Codon-n-
10 Tupel-Häufigkeit des resultierenden zweiten Codon-n-Tupels. Vorzugsweise umfasst das Verfahren weiterhin einen Schritt d) Bewerten der in Schritt c) ermittelten relativen Codon-n-Tupel-Häufigkeit des zweiten Codon-n-Tupels, wobei die Bewertung z. B. ein Vergleichen mit der relativen Codon-n-Tupel-Häufigkeit des
15 ersten Codon-n-Tupels beinhalten und/oder anhand eines Zielkriteriums wie beispielsweise einem Mindestwert oder dergleichen durchgeführt werden kann. Diesbezüglich wird auch auf die nachfolgende Beschreibung und die Ausführungsbeispiele verwiesen. Die Schritte b) und c) sowie ggf. d) können auch
20 wiederholt werden, bis ein zweites Codon-n-Tupel resultiert, das die erfindungsgemäß erforderliche höhere relative Codon-n-Tupel-Häufigkeit aufweist als das erste Codon-n-Tupel.

Wie bereits oben dargestellt wurde, sind der ermittelten Menge
25 der Aminosäure-n-Tupel-Ereignisse bzw. der ermittelten relativen Häufigkeit des Codon-n-Tupels nicht notwendigerweise das vollständige Genom des mindestens einen Zielorganismus oder die vollständigen Genome der zur Infektion des mindestens einen Zielorganismus befähigten Viren zugrunde zu legen. Vielmehr kann
30 es in bestimmten Verfahrensvarianten ausreichend und vorteilhaft sein, wenn lediglich Teile des Genoms bzw. der Genome von der ermittelten absoluten Aminosäure-n-Tupel-Häufigkeit bzw. der relativen Codon-n-Tupel-Häufigkeit umfasst sind, zumal die

meisten Genome jeweils zu einem Großteil nichtcodierende Bereiche enthalten, die für die erfindungsgemäße Bemessung der Eignung einer vorgegebenen Nukleinsäuresequenz für die Expression in dem Zielorganismus weniger relevant sind.

5

In bevorzugten Ausführungsformen ergibt sich die Menge von Aminosäure-n-Tupel-Ereignissen aus mehreren, vorzugsweise sämtlichen, proteincodierenden Genen und/oder Proteinen des mindestens einen Zielorganismus bzw. der zur Infektion des
10 mindestens einen Zielorganismus befähigten Viren bzw. wird anhand von mehreren proteincodierenden Genen und/oder Proteinen des mindestens einen Zielorganismus bzw. der zur Infektion des mindestens einen Zielorganismus befähigten Viren bestimmt. Konstitutiv vom Zielorganismus exprimierte Proteine oder
15 Proteine mit hoher transienter Expression bzw. hoher Abundanz sind hierfür besonders geeignet. Besonders bevorzugt sind mindestens 25%, mindestens 50%, mindestens 75%, mindestens 80%, mindestens 90% oder mindestens 95% vom codierenden Teil des Genoms des Zielorganismus von der Ermittlung der Menge von
20 Aminosäure-n-Tupel-Ereignissen umfasst.

Entsprechend ergibt sich auch die relative Codon-n-Tupel-Häufigkeit aus einer Menge von Ereignissen des jeweils ersten bzw. zweiten Codon-n-Tupels in mehreren, vorzugsweise
25 sämtlichen, proteincodierenden Genen und/oder Proteinen des mindestens einen Zielorganismus bzw. der zur Infektion des mindestens einen Zielorganismus befähigten Viren bzw. wird anhand von mehreren proteincodierenden Genen und/oder Proteinen des mindestens einen Zielorganismus bzw. der zur Infektion des
30 mindestens einen Zielorganismus befähigten Viren bestimmt, der die Menge von Aminosäure-n-Tupel Ereignissen zugrunde gelegt wird. Besonders bevorzugt sind mindestens 25%, mindestens 50%, mindestens 75%, mindestens 80%, mindestens 90% oder mindestens

95% vom codierenden Teil des Genoms des Zielorganismus von der Ermittlung der relativen Codon-n-Tupel-Häufigkeit umfasst.

Es ist für das erfindungsgemäße Verfahren grundsätzlich nicht
5 erforderlich, dass die relative Häufigkeit des ersten bzw.
zweiten Codon-n-Tupels ein aus absoluten Häufigkeiten
arithmetisch errechneter Wert ist. Beispielsweise kann die
relative Häufigkeit des ersten und/oder zweiten Codon-n-Tupels
auch durch eine andere Wahrscheinlichkeitsverteilung bzw. ein
10 anderes Wahrscheinlichkeitsmaß ausgedrückt oder ersetzt werden,
z. B. als Intervallschätzung. So ist es z. B. möglich, aus
zufälligen Beobachtungen, z. B. in Form von Stichproben im Genom
des mindestens einen Zielorganismus oder eines Teils davon bzw.
in den Genomen der zur Infektion des mindestens einen
15 Zielorganismus befähigten Viren oder der Teile davon, mit der
beobachteten ersten bzw. zweiten Codon-n-Tupel-Häufigkeit
und/oder der beobachteten Häufigkeit sämtlicher synonymen Codon-
n-Tupel, die für das gleiche Aminosäure-n-Tupel codieren, als
Anzahl von Erfolgen und der Aminosäure-n-Tupel-Häufigkeit als
20 Anzahl von Versuchen ein Konfidenzintervall für die
Erfolgswahrscheinlichkeit einer Binomial- oder auch
Multinomialverteilung zu bestimmen, beispielsweise das Clopper-
Pearson-Vertrauensintervall mit einem Konfidenzniveau von 95%
oder simultane Konfidenzintervalle für multinomiale Anteile. Die
25 relative Codon-n-Tupel-Häufigkeit wird dann z. B. durch den
Intervallmittelpunkt repräsentiert bzw. ersetzt, welcher eine
genauere Schätzung darstellt. Dies kann insbesondere bei
geringer Aminosäure-n-Tupel-Häufigkeit und bei sehr hoher oder
sehr niedriger relativer Codon-n-Tupel-Häufigkeit vorteilhaft
30 sein. Selbstverständlich kann die relative Codon-n-Tupel-
Häufigkeit aber auch mithilfe anderer Werte aus dem Intervall
berechnet bzw. repräsentiert werden. Außer dem
Intervallmittelpunkt kann z. B. der kleinste Wert des

Intervalls, insbesondere für sehr konservative Schätzungen, oder der Durchschnittswert einer Gewichtsfunktion (z. B. $-1/x$) auf dem Intervall die relative Codon-n-Tupel-Häufigkeit repräsentieren.

5

Eine Variante des erfindungsgemäßen Verfahrens sieht daher vor, dass die mindestens eine Änderungsposition eine direkte Aufeinanderfolge von n Basentriplets umfasst, die ein erstes Codon-n-Tupel bildet und einen Sequenzabschnitt von n Aminosäuren der vorgegebenen Aminosäuresequenz codiert, der ein Aminosäure-n-Tupel bildet, wobei mindestens eines der n Basentriplets der direkten Aufeinanderfolge durch ein synonymes Basentriplett ersetzt wird, welches anhand einer Schätzfunktion so ausgewählt wird, dass ein zweites Codon-n-Tupel resultiert, welches das Aminosäure-n-Tupel mit einer größeren Wahrscheinlichkeit im Genom oder eines Teils davon des mindestens einen Zielorganismus und/oder in Genomen oder Teilen davon von zur Infektion des mindestens einen Zielorganismus befähigten Viren codiert als das erste Codon-n-Tupel. Geeignete Schätzfunktionen, welche für das erfindungsgemäße Verfahren implementiert werden können, sind dem Fachmann bekannt.

So, wie der Begriff hier verwendet wird, schließt „mindestens eine Änderungsposition“ die Möglichkeit mit ein, dass das Verfahren mehrere bzw. eine Vielzahl von Änderungspositionen umfasst, wobei in jeder der Änderungspositionen mindestens eines der n Basentriplets der direkten Aufeinanderfolge durch ein synonymes Basentriplett ersetzt wird und die synonymen Basentriplets hierbei so gewählt werden, dass zumindest ein Teil der resultierenden zweiten Codon-n-Tupel eine höhere relative Codon-n-Tupel-Häufigkeit aufweist als die jeweiligen ersten Codon-n-Tupel. Selbstverständlich können zu diesem Zweck in einer oder mehrerer der Änderungspositionen auch mehrere oder

sämtliche der n Basentriplets jeweils durch ein synonymes Basentriplett ersetzt werden.

In bevorzugten Ausführungsformen wird hierbei zumindest in der
5 Änderungsposition mit dem ersten Codon-n-Tupel, das von allen
Änderungspositionen die geringste relative Codon-n-Tupel-
Häufigkeit aufweist, mindestens eines der Basentriplets der
direkten Aufeinanderfolge durch ein synonymes Basentriplett
ersetzt, welches so gewählt ist, dass das resultierende zweite
10 Codon-n-Tupel eine höhere relative Codon-n-Tupel Häufigkeit als
das erste Codon-n-Tupel aufweist. Die Erfinder haben erkannt,
dass sich das Codon-n-Tupel mit der geringsten relativen
Häufigkeit in der Nukleotidsequenz regelmäßig limitierend auf
den gesamten Translationsprozess auswirkt, sodass der Austausch
15 dieses ersten Codon-n-Tupels durch ein zweites Codon-n-Tupel mit
höherer relativer Häufigkeit sich besonders vorteilhaft auf die
Translation und Faltung auswirken und dadurch zu einer besonders
starken Verbesserung der Expression löslichen Proteins im
Zielorganismus führen kann.

20

Bei mehreren oder einer Vielzahl von Änderungspositionen in der
Nukleotidsequenz sieht das erfindungsgemäße Verfahren die
Möglichkeit vor, dass sich die direkten Aufeinanderfolgen der n
Basentriplets von mindestens zwei Änderungspositionen
25 überschneiden, wobei das Basentriplett, welches durch das
synonyme Basentriplett ersetzt wird, von mindestens diesen zwei
Änderungspositionen gleichzeitig umfasst ist.

Für solche Ausführungsformen kann das erfindungsgemäße Verfahren
30 weiterhin vorsehen, dass das synonyme Basentriplett so gewählt
wird, dass in einer der beiden Änderungspositionen das
resultierende zweite Codon-n-Tupel eine geringere relative
Codon-n-Tupel Häufigkeit und in der anderen der beiden

Änderungspositionen das resultierende zweite Codon-n-Tupel eine höhere relative Codon-n-Tupel-Häufigkeit als das jeweilige erste Codon-n-Tupel aufweist. Alternativ oder zusätzlich kann das erfindungsgemäße Verfahren aber für solche sich überschneidenden
5 Änderungspositionen auch vorsehen, dass das resultierende zweite Codon-n-Tupel in beiden Änderungspositionen eine höhere relative Codon-n-Tupel-Häufigkeit als das jeweilige erste Codon-n-Tupel aufweist. Es ist weiterhin möglich, dass sich die relative Codon-n-Tupel-Häufigkeit in beiden Änderungspositionen
10 verringert.

Wenn für n ein Wert größer gleich zwei gewählt wird, ist es selbstverständlich auch möglich, dass sich die direkten Aufeinanderfolgen der n Basentriplets von mehr als zwei
15 Änderungspositionen überschneiden und das Basentriplett, welches durch das synonyme Basentriplett ersetzt wird, von mehr als zwei Änderungspositionen gleichzeitig umfasst ist. Es ist dann möglich, dass das synonyme Basentriplett so gewählt wird, dass in mindestens einer der Änderungspositionen das resultierende
20 zweite Codon-n-Tupel eine geringere relative Codon-n-Tupel Häufigkeit und in den anderen der Änderungspositionen das resultierende zweite Codon-n-Tupel eine höhere relative Codon-n-Tupel-Häufigkeit als das jeweilige erste Codon-n-Tupel aufweist.

25 Es hat sich bei dem erfindungsgemäßen Verfahren überraschenderweise gezeigt, dass bei einer Vielzahl von Änderungspositionen in der Nukleotidsequenz, die sich zumindest teilweise überschneiden, die Inkaufnahme einer Verschlechterung, d. h. eine Verringerung der relativen Codon-n-Tupel-Häufigkeit
30 eines zweiten Codon-n-Tupels gegenüber dem erste Codon-n-Tupel, von einzelnen Codon-n-Tupeln erforderlich und sogar wesentlich sein kann, um insgesamt eine Verbesserung der Translationseffizienz der gesamten Nukleotidsequenz zu

erreichen. Das kann insbesondere der Fall sein, wenn sich eine Änderungsposition mit einem für die Expression in dem Zielorganismus besonders ungünstigen ersten Codon-n-Tupel mit einer Änderungsposition mit einem für die Expression in den
5 Zielorganismus günstigen ersten Codon-n-Tupel überschneidet und kein synonymes Basentriplett gewählt werden kann, dass die relative Häufigkeit für beide zweite Codon-n-Tupel steigert.

Es ist z. B. möglich, dass die relative Codon-n-Tupel-Häufigkeit der zweiten Codon-n-Tupel gegenüber den jeweiligen ersten Codon-n-Tupeln in mindestens etwa 1%, 5% oder 10% und/oder höchstens etwa 40%, 30% oder 20% der Änderungspositionen verringert wird.
10

In diesem Zusammenhang haben die Erfinder weiterhin erkannt, dass es für die Verbesserung der Expressionsrate löslichen Proteins aus der Nukleotidsequenz in dem Zielorganismus wesentlich wichtiger sein kann, mithilfe des erfindungsgemäßen Verfahrens besonders niedrige relative Codon-n-Tupel-Häufigkeiten zu erhöhen als vereinzelt oder auch mehrheitlich
15 besonders hohe relative Codon-n-Tupel Häufigkeiten zu erreichen. Das erfindungsgemäße Verfahren sieht daher in bevorzugten Ausführungsformen als Maßgabe bzw. Ziel für die Optimierung vor, dass die relative Codon-n-Tupel-Häufigkeit der ersten Codon-n-Tupel und die relative Codon-n-Tupel-Häufigkeit der zweiten
20 Codon-n-Tupel in den Änderungspositionen jeweils ein globales Minimum aufweist, wobei das globale Minimum der zweiten Codon-n-Tupel größer ist als das globale Minimum der ersten Codon-n-Tupel.
25

In weiter bevorzugten Ausführungsformen werden die synonymen Basentriplets so gewählt, dass die relative Codon-n-Tupel Häufigkeit der zweiten Codon-n-Tupel einen größtmöglichen Mindestwert, d. h. ein größtmögliches globales Minimum, erreicht
30

oder zumindest nicht mehr als 50%, vorzugsweise nicht mehr als 40% oder nicht mehr als 30%, bevorzugt nicht mehr als 20%, besonders bevorzugt nicht mehr als 10% unterhalb des größtmöglichen Mindestwertes liegt. Geeignete mathematische

5 Näherungs- und/oder Optimierungsverfahren zum Auffinden eines größtmöglichen globalen Minimums sind dem Fachmann bekannt. Auch hier liegt eine Abkehr von der bisher vorherrschenden technischen Lehre aus Codonoptimierung bzw. Codonkontextoptimierung vor, indem das erfindungsgemäße

10 Verfahren vorrangig auf eine globale Optimierung von ungünstigen Änderungspositionen in der Nukleotidsequenz abstellt, die sich erfindungsgemäß durch geringe Codon-n-Tupel-Häufigkeiten identifizieren lassen, und nicht auf das Erreichen lokaler Höchstwerte in Bezug auf Codonverwendung oder

15 Codonpaarpräferenz. Auf diese Weise lassen sich insbesondere auch Proteine, die mithilfe von herkömmlichen Optimierungsverfahren bisher kaum oder gar nicht in artfremden Zielorganismen exprimiert werden konnten, sehr zuverlässig für eine heterologe Expression zugänglich machen.

20 Vorzugsweise werden die synonymen Basentriplets so gewählt, dass ein Mittelwert aus den relativen Codon-n-Tupel-Häufigkeiten der zweiten Codon-n-Tupel einen Höchstwert erreicht oder zumindest nicht mehr als 50%, vorzugsweise nicht mehr als 40%

25 oder nicht mehr als 30%, bevorzugt nicht mehr als 20%, besonders bevorzugt nicht mehr als 10% unterhalb eines erreichbaren Höchstwertes liegt. Auf diese Weise lassen sich mithilfe des Optimierungsverfahrens Nukleotidsequenzen bereitstellen, die besonders gut an die Expression in dem Zielorganismus angepasst

30 sind und sich dementsprechend besonders zuverlässig und mit hohen Expressionsraten löslichen Proteins in dem Zielorganismus exprimieren lassen.

Es versteht sich, dass für die vorgenannten Ausführungsformen eine Ermittlung der relativen Häufigkeit der ersten Codon-n-Tupel nicht erforderlich ist, da das Erreichen der genannten Optimierungskriterien bereits eine höhere relative Codon-n-Tupel Häufigkeit von zumindest einem Teil der zweiten Codon-n-Tupel im Vergleich zu den jeweiligen ersten Codon-n-Tupeln impliziert. Mit anderen Worten ist die erfindungsgemäße Grundbedingung, dass mindestens eines der n Basentriplets aus der direkten Aufeinanderfolge der mindestens einen Änderungsposition durch ein synonymes Basentriplett ersetzt wird, das gezielt so gewählt ist, dass mindestens ein zweites Codon-n-Tupel resultiert, das in Bezug auf die Menge der Aminosäure-n-Tupel-Ereignisse eine höhere relative Codon-n-Tupel Häufigkeit in dem Genom oder des Teils davon des mindestens einen Zielorganismus und/oder in den Genomen oder den Teilen davon der zur Infektion des mindestens einen Zielorganismus befähigten Viren aufweist als das erste Codon-n-Tupel, beim Erreichen der genannten Kriterien zwangsläufig immer erfüllt. Die relative Häufigkeit der ersten Codon-n-Tupel kann aber grundsätzlich als Kontrollparameter für die Verwirklichung des erfindungsgemäßen Verfahrens hinzugezogen werden oder beispielsweise um ein Zwischen- oder Endergebnis der Optimierung mit dem Ausgangszustand zu vergleichen. Im Übrigen kann in diesen Ausführungsformen die relative Häufigkeit der ersten Codon-n-Tupel dahingestellt bleiben, da sich das Ziel der Optimierung nicht daran bemisst, mit welcher Ausgangssequenz gestartet wurde.

Wie bereits im Vorhergehenden mehrfach dargelegt wurde, haben die Erfinder erkannt, dass es für die Optimierung einer Nukleotidsequenz oft vorteilhafter ist, den Schwerpunkt des Verfahrens stärker auf die Optimierung von Änderungspositionen mit niedriger relativer Codon-n-Tupel-Häufigkeit zu legen als auf eine Erhöhung von durchschnittlichen oder ohnehin bereits

hohen relativen Codon-n-Tupel-Häufigkeiten. Zu diesem Zweck ist in bevorzugten Ausführungsformen des erfindungsgemäßen Verfahrens vorgesehen, dass der Mittelwert eine degressive Wichtung der relativen Codon-n-Tupel-Häufigkeiten der ersten und
5 zweiten Codon-n-Tupel enthält, die so konfiguriert ist, dass eine hohe relative Codon-n-Tupel-Häufigkeit im Vergleich zu einer niedrigeren relativen Codon-n-Tupel Häufigkeit einen betragsmäßig unterproportionalen Einfluss auf den Mittelwert bzw. dessen Berechnung hat. Umgekehrt kann sich dadurch die
10 Erhöhung einer niedrigen relativen Codon-n-Tupel-Häufigkeit, auch wenn diese geringfügig ist, stärker im Mittelwert niederschlagen als die Erhöhung einer mittleren oder hohen relativen Codon-n-Tupel-Häufigkeit.

15 Selbstverständlich kann das Verfahren aber zusätzlich oder alternativ auch darauf ausgelegt sein, dass die relative Codon-n-Tupel-Häufigkeit in zumindest einem Teil der Änderungspositionen jeweils einen Höchstwert erreicht.

20 Es versteht sich, dass die natürliche Zahl n für die n Basentriplets, das Codon- n -Tupel und das Aminosäure- n -Tupel einer Änderungsposition die gleiche natürliche Zahl sein muss. Wenn die Änderungsposition z. B. eine direkte Aufeinanderfolge von drei Basentriplets umfasst, codiert diese auch für einen
25 Sequenzabschnitt von drei Aminosäuren der vorgegebenen Aminosäuresequenz, d. h. das erste und zweite Codon- n -Tupel sind jeweils ein Codon-3-Tupel und das Aminosäure- n -Tupel ist entsprechend ein Aminosäure-3-Tupel. Dessen ungeachtet ist es aber möglich, dass sich verschiedene Änderungspositionen, z. B.
30 mindestens zwei Änderungspositionen, in der Zahl n unterscheiden bzw. dass n für verschiedene Änderungspositionen, z. B. für mindestens zwei Änderungspositionen, unterschiedlich gewählt wird. Darüber hinaus ist vorgesehen, dass das n einer

Änderungsposition während des Verfahrens variiert werden kann, d. h. die Änderungsposition kann während des Verfahrens vergrößert oder verkleinert bzw. eine große Änderungsposition in mehrere kleine Änderungspositionen unterteilt werden und
5 umgekehrt. Das kann z. B. vorteilhaft sein, um in Bereichen der Nukleotidsequenz, die Aminosäure-n-Tupel codieren, die selten oder gar nicht im Genom des Zielorganismus oder der Zielorganismen oder Viren vorkommen bzw. in den Teilen davon, auf kleinere n mit sichererer Statistik zurückzugreifen.
10 Beispielsweise kann es vorteilhaft sein, einen Abschnitt der Nukleotidsequenz mit drei Codons durch zwei überlappende Änderungspositionen mit $n = 2$ zu optimieren als durch eine Änderungsposition mit $n = 3$, wenn das zugehörige Aminosäure-3-Tupel nur selten oder, im Extremfall, gar nicht vorkommt. Ein
15 weiterer Vorteil ist, dass das erfindungsgemäße Verfahren Nukleotidabfolgen, die für die Expression in dem Zielorganismus schädlich sind, wie z. B. Restriktionsschnittstellen, durch die Verwendung größerer n mit großer Wahrscheinlichkeit ausschließt. Da solche Sequenzmotive in der Regel keine Grundlage im eigenen
20 Genom des Zielorganismus haben, d. h. die relative Codon-n-Tupel-Häufigkeit im Genom des Zielorganismus oder der Zielorganismen geht gegen Null, führt das erfindungsgemäße Verfahren implizit zu deren systematischem Ausschluss. Solche schädlichen Sequenzen werden z. B. bis zu einer Länge von
25 einschließlich $3n-2$ bereits bei einem n größer gleich 3 automatisch aus der optimierten Nukleotidsequenz ausgeschlossen.

In bevorzugten Ausführungsformen wird das Ersetzen der Basentriplets durch die synonymen Basentriplets in mehreren
30 Iterationsschritten durch ein computergestütztes Optimierungsverfahren durchgeführt. Auf diese Weise ist es insbesondere möglich, Nukleotidsequenzen mit einer Vielzahl von einander überlappenden Änderungspositionen sukzessive zu

optimieren. In einem Iterationsschritt kann in allen
Änderungspositionen das mindestens eine der n Basentriplets
durch ein synonymes Basentriplett ersetzt werden oder nur in
einem Teil der Änderungspositionen. Vorzugsweise wird in jedem
5 Iterationsschritt nur in einem Teil der Änderungspositionen das
mindestens eine der n Basentriplets durch ein synonymes
Basentriplett ersetzt. Bevorzugt wird zumindest in einem Teil
der Iterationsschritte, besonders bevorzugt im überwiegenden
Teil oder jedem der Iterationsschritte jeweils nur ein
10 Basentriplett durch ein synonymes Basentriplett ersetzt. Das
Basentriplett kann dabei von einer Änderungsposition oder
mehreren sich überschneidenden Änderungspositionen umfasst sein.
Das Ersetzen der Basentriplets durch die synonymen
Basentriplets kann beispielsweise so oft iteriert werden, bis
15 eines der bereits oben genannten Zielkriterien erreicht wird,
beispielsweise also die relativen Codon-n-Tupel-Häufigkeiten der
zweiten Codon-n-Tupel einen größtmöglichen Mindestwert, einen
größtmöglichen Mittelwert, insbesondere einen größtmöglichen
gewichteten Mittelwert, oder einen Höchstwert erreichen oder in
20 den oben definierten Maßen an diese Werte heranreicht.

Es versteht sich, dass das Verfahren in diesen Ausführungsformen
auch einen oder mehrere Iterationsschritte umfassen kann, die
von dem jeweiligen Zielkriterium wegführen, insbesondere indem
25 die relative Häufigkeit des zweiten Codon-n-Tupels in einer oder
mehreren der Änderungspositionen gegenüber dem ersten Codon-n-
Tupel zumindest zwischenzeitlich durch einen Iterationsschritt
erniedrigt wird, um lokale Maxima von Mindestwert bzw.
Mittelwert überwinden zu können, die dem Erreichen des
30 Zielkriteriums entgegenstehen. Es ist daher auch vorgesehen,
dass in einer oder mehreren der Änderungspositionen das
mindestens eine der n Basentriplets bis zum Erreichen des
Zielkriteriums auch mehrfach durch unterschiedliche synonyme

Basentripletts ersetzt werden kann. Insbesondere ist eine Festlegung einer Änderungsposition auf ein bestimmtes zweites Codon-n-Tupel nach Durchführung eines Iterationsschrittes insoweit nicht vorgesehen.

5

Vorzugsweise umfasst das computergestützte Optimierungsverfahren ein Approximationsverfahren, insbesondere ein simuliertes Abkühlungsverfahren, auch als „Simulated Annealing“ bezeichnet. Diese Verfahren haben sich zum Auffinden einer Näherungslösung für eine bezüglich der relativen Codon-n-Tupel-Häufigkeiten für den Zielorganismus optimalen Nukleotidsequenz als besonders geeignet und vorteilhaft erwiesen, insbesondere wenn längere Nukleotidsequenzen, z. B. mit 30 Codons oder mehr, mit einer Vielzahl überlappender Änderungspositionen aufgrund ihrer Komplexität das vollständige Überprüfen von allen möglichen synonymen Basentripletts und mathematische Optimierungsverfahren ausschließen. Selbstverständlich sind aber auch andere heuristische Approximationsverfahren wie z. B. ein Sintflutalgorithmus oder ein genetischer Algorithmus möglich. Weitere geeignete Approximationsverfahren sind dem Fachmann bekannt. Darüber hinaus sind auch andere computergestützte Optimierungsverfahren wie z. B. künstliche Intelligenz (KI)-basierte Anwendungen denkbar.

25 In bevorzugten Ausführungsformen des Verfahrens umfassen die Änderungspositionen zusammen mindestens 1%, mindestens 5%, mindestens 10%, mindestens 20%, mindestens 30%, mindestens 40% oder mindestens 50% der Basentripletts der Nukleotidsequenz, die eine Aminosäure der vorgegebenen Aminosäuresequenz codieren.

30 Vorzugsweise umfassen die Änderungspositionen zusammen mindestens 60%, mindestens 70% oder mindestens 80%, besonders bevorzugt mindestens 90% oder mindestens 95% der Basentripletts der Nukleotidsequenz, die eine Aminosäure der vorgegebenen

Aminosäuresequenz codieren. Auf diese Weise gewährleistet das erfindungsgemäße Verfahren eine besonders zuverlässige Optimierung der Nukleotidsequenz für die Expression in dem Zielorganismus. Wie oben erwähnt können Änderungspositionen auch

5 Basentriplets enthalten, die nicht durch ein synonymes Basentriplett ersetzt werden, d. h. es ist nicht erforderlich, dass jedes von den Änderungspositionen umfasste Basentriplett durch ein synonymes Basentriplett ersetzt wird.

10 Vorzugsweise werden die synonymen Basentriplets so gewählt, dass die relativen Codon-n-Tupel-Häufigkeiten der zweiten Codon-n-Tupel in mindestens 10%, mindestens 15%, mindestens 20%, mindestens 25%, mindestens 30% oder mindestens 35%, vorzugsweise in mindestens 40% oder mindestens 45% der Änderungspositionen,

15 bevorzugt in mindestens 50%, mindestens 55%, mindestens 60%, mindestens 65% oder mindestens 70% der Änderungspositionen, besonders bevorzugt in mindestens 75%, mindestens 80%, mindestens 85% oder mindestens 90% der Änderungspositionen eine höhere relative Codon-n-Tupel-Häufigkeit aufweisen als die

20 jeweiligen ersten Codon-n-Tupel.

So, wie der Begriff hier verwendet wird, schließt „mindestens ein Zielorganismus“ die Möglichkeit mit ein, dass die Nukleotidsequenz für die Expression der vorgegebenen

25 Aminosäuresequenz in einer Mehrzahl von verschiedenen Zielorganismen gleichzeitig optimiert wird. Das erfindungsgemäße Verfahren sieht dafür z. B. vor, dass sich die relative Codon-n-Tupel-Häufigkeit aus der Menge von Ereignissen eines Codon-n-Tupels bezogen auf die Menge von Ereignissen des entsprechenden

30 Aminosäure-n-Tupels in den Genomen oder den Teilen davon der Mehrzahl der verschiedenen Zielorganismen ergibt. Auf diese Weise kann mithilfe des erfindungsgemäßen Verfahrens eine optimierte Nukleotidsequenz bereitgestellt werden, die sich

besser für die heterologe Expression in verschiedenen Expressionssystemen eignet. Dies ist mit herkömmlichen Methoden, die z. B. auf Codon Usage oder Codon Adaptation Index basieren, nicht möglich.

5

Der mindestens eine Zielorganismus kann grundsätzlich eine vorbestimmte Wirtszelle oder ein beliebiger vorbestimmter Organismus sein, die bzw. der für die Expression der vorgegebenen Aminosäuresequenz geeignet ist. Bei der Wirtszelle
10 kann es sich um eine prokaryotische oder eine eukaryotische Wirtszelle handeln. Bei der Wirtszelle kann es sich um eine Wirtszelle handeln, die für die Kultur in flüssigen oder festen Medien geeignet ist. Alternativ kann es sich bei der Wirtszelle um eine Zelle handeln, die Teil eines multizellulären Gewebes
15 oder eines multizellulären Organismus wie einer, insbesondere transgenen, Pflanze, eines Tieres oder eines Menschen ist.

Die Wirtszelle kann mikrobiell oder nicht-mikrobiell sein. Eine mikrobielle Wirtszelle kann eine Bakterien-, Hefe- oder
20 Pilzzelle sein. Geeignete bakterielle Wirtszellen sind sowohl Gram-positive als auch Gram-negative Bakterien. Beispiele für geeignete bakterielle Wirtszellen sind Bakterien aus den Gattungen *Bacillus*, *Actinomycetis*, *Escherichia*, *Streptomyces* sowie Milchsäurebakterien wie *Lactobacillus*, *Streptococcus*,
25 *Lactococcus*, *Oenococcus*, *Leuconostoc*, *Pediococcus*, *Carnobacterium*, *Propionibacterium*, *Enterococcus* und *Bifidobacterium*. Besonders bevorzugt sind *Bacillus subtilis*, *Bacillus amyloliquefaciens*, *Bacillus licheniformis*, *Escherichia coli*, *Streptomyces coelicolor*, *Streptomyces clavuligerus*,
30 *Lactobacillus plantarum* und *Lactococcus lactis*, insbesondere *Escherichia coli*.

- Alternativ kann die Wirtszelle auch ein eukaryontischer Mikroorganismus wie eine Hefe oder ein, insbesondere filamentöser, Pilz sein. Bevorzugte Hefen als Wirtszellen gehören zu den Gattungen *Saccharomyces*, *Kluyveromyces*, *Candida*,
5 *Pichia*, *Schizosaccharomyces*, *Hansenula*, *Kloeckera*, *Schwanniomyces*, und *Yarrowia*. Besonders bevorzugte *Debaromyces*-Wirtszellen sind *Saccharomyces cerevisiae* und *Kluyveromyces lactis*.
- 10 Gemäß einer weiteren bevorzugten Ausführungsform ist die Wirtszelle der vorliegenden Erfindung eine Zelle eines filamentösen Pilzes. Filamentöse Pilze umfassen alle fadenförmigen Formen der Unterabteilung Eumycota und Oomycota. Filamentöse Pilze zeichnen sich durch eine Myzelwand aus, die
15 aus Chitin, Cellulose, Glucan, Chitosan, Mannan und anderen komplexen Polysacchariden besteht. Das vegetative Wachstum erfolgt durch Hyphenstreckung und der Kohlenstoffabbau ist obligatorisch aerob. Zu den filamentösen Pilzen, deren Stämme als Wirtszellen in der vorliegenden Erfindung verwendet werden
20 können, gehören unter anderem Stämme der Gattungen *Acremonium*, *Aspergillus*, *Aureobasidium*, *Cryptococcus*, *Filibasidium*, *Fusarium*, *Humicola*, *Magnaporthe*, *Mucor*, *Myceliophthora*, *Neocallimastix*, *Neurospora*, *Paecilomyces*, *Penicillium*, *Piromyces*, *Schizophyllum*, *Chrysosporium*, *Talaromyces*, *Thermoascus*, *Thielavia*,
25 *Tolypocladium* und *Trichoderma*. Bevorzugte Arten filamentöser Pilze sind ausgewählt aus der Gruppe bestehend aus *Aspergillus niger*, *Aspergillus oryzae*, *Aspergillus sojae*, *Trichoderma reesei* und *Penicillium chrysogenum*. Beispiele für geeignete Wirtsstämme sind dem Fachmann bekannt.
- 30 Geeignete nicht-mikrobielle Wirtszellen sind beispielsweise Säugetierwirtszellen wie Hamsterzellen (z. B. Chinese Hamster Ovary (CHO)-Zellen; Baby Hamster Kidney (BHK)-Zellen),

Mauszellen, Affenzellen oder menschliche Zellen oder Zelllinien wie HeLa oder HEK293; Insektenzellen wie Drosophila-Zellen oder Lepidoptera-Zelllinien Hi5, Sf21; Pflanzenzellen wie z. B. Zellen von Tabak, Tomate, Kartoffel, Raps, Kohl, Erbse, Weizen, 5 Mais, Reis, Taxus-Arten wie *Taxus brevifolia*, Arabidopsis-Arten wie *Arabidopsis thaliana*, und Nicotiana-Arten wie *Nicotiana tabacum*. Außerdem sind nicht-pathogene Leishmanien für die Proteinexpression geeignet. Solche nicht-mikrobiellen Zellen eignen sich besonders für die Herstellung von Säugetier- oder 10 Humanproteinen zur Verwendung in der Säugetier- oder Humantherapie.

Die vorgegebene Aminosäuresequenz ist vorzugsweise ein Protein oder ein Teil davon, das insbesondere natürlicherweise ein 15 eukaryotisches Protein ist. Das erfindungsgemäße Verfahren hat sich als besonders vorteilhaft für die Expression eukaryotischer Proteine, beispielsweise eines Insekten-, Pflanzen- oder Säugerproteins, in bakteriellen, insbesondere prokaryotischen, Expressionssystemen wie beispielsweise *Escherichia coli* 20 erwiesen.

In Verfahrensvarianten, in denen die Nukleotidsequenz für die Expression der vorgegebenen Aminosäuresequenz in einer Mehrzahl von verschiedenen Zielorganismen gleichzeitig optimiert wird, 25 ist es möglich, dass die verschiedenen Zielorganismen sehr unterschiedliche Genomgrößen aufweisen, wobei große Genome in der Regel Proteome mit einer größeren Anzahl von Aminosäure-n-Tupeln codieren als kleinere Genome. Das kann dazu führen, dass die relativen Codon-n-Tupel-Häufigkeiten überproportional durch 30 große Genome beeinflusst werden und somit die Optimierung der Nukleotidsequenz zwangsläufig eine Expression in Zielorganismen mit großen Genomen stärker begünstigt als in Zielorganismen mit kleinen Genomen. Für solche Fälle sieht das erfindungsgemäße

Verfahren vor, dass die relative Codon-n-Tupel Häufigkeit eine genomabhängige Wichtung enthält, die beispielsweise so konfiguriert ist, dass ein Größenunterschied in den Genomen oder den Teilen davon, insbesondere ein unterschiedlicher Umfang der codierenden Bereiche, der verschiedenen Zielorganismen zumindest teilweise kompensiert wird. Auf diese Weise wird gewährleistet, dass sich die erfindungsgemäß optimierte Nukleotidsequenz besser für die Expression in den verschiedenen Zielorganismen eignet.

Überraschenderweise hat sich gezeigt, dass das erfindungsgemäße Verfahren nicht nur in der Lage ist, die mengenmäßige Proteinausbeute im heterologen Expressionssystem gegenüber herkömmlichen Optimierungsverfahren deutlich zu verbessern, sondern insbesondere auch den Anteil löslichen Proteins in der Ausbeute signifikant zu steigern. Hierin liegt ein besonderer Vorteil des erfindungsgemäßen Verfahrens, da in der Regel die gelöste Form eines Proteins den nativen, biochemisch aktiven Zustand darstellt, der vor allem bei Hochwertproteinen für wissenschaftliche, medizinisch-pharmazeutische und biotechnologische Zwecke von zentraler Bedeutung ist. Ein besonderes Kennzeichen des erfindungsgemäßen Verfahrens ist somit, dass nach der Optimierung der Nukleotidsequenz die in dem mindestens einen Zielorganismus exprimierte Aminosäuresequenz eine größere Löslichkeit aufweist und/oder zu einem größeren Anteil in gelöster Form vorliegt als vor der Optimierung.

Wie sich bereits aus der obigen Beschreibung ergibt, kann die Optimierung der Nukleotidsequenz für die Expression in dem mindestens einen Zielorganismus alternativ oder zusätzlich zur Optimierung anhand des Genoms des Zielorganismus oder Teilen davon auch anhand von Genomen oder Teilen davon von zur Infektion des mindestens einen Zielorganismus befähigten Viren erfolgen. Insbesondere können die Viren Bakteriophagen umfassen.

Hierfür ist es wesentlich, dass die Optimierung anhand einer Mehrzahl oder Vielzahl von Virengenomen bzw. Teilen davon erfolgt, da ein einzelnes Virengenom bzw. Virentranskriptom in
5 der Regel nicht den erforderlichen Umfang und somit nicht die nötige statistische Signifikanz für eine wirksame Optimierung der Nukleotidsequenz anhand der erfindungsgemäßen relativen Codon-n-Tupel Häufigkeit hat. Die Erfinder fassen daher die Genome bzw. Transkriptome verschiedener Viren zu einer Art
10 „Supergenom“ bzw. „Supertranskriptom“ zusammen, das der Ermittlung der relativen Codon-n-Tupel-Häufigkeit zugrunde gelegt wird. Vorzugsweise umfasst das Verfahren in diesen Ausführungsformen mindestens zehn, mindestens 20, mindestens 30, mindestens 40, mindestens 50, mindestens 100, mindestens 150
15 oder mindestens 200 Genome verschiedener Viren oder entsprechende Teile davon.

Die Erfinder machen sich hierbei die Erkenntnis zu Nutze, dass die Genome von Viren bzw. Phagen bereits von Natur aus auf eine
20 Proteinexpression im Hochdurchsatz in dem infizierten Zielorganismus optimiert sind. In diesem Zusammenhang haben die Erfinder als besonderen Vorteil erkannt, dass Viren- bzw. Phagengenome oft eine reduzierte Basenkomplexität aufweisen, die dazu führt, dass eine mRNA, die von einer erfindungsgemäß
25 optimierten Nukleotidsequenz transkribiert wird, kaum oder keine Sekundärstrukturen ausbildet. Dadurch ist das erfindungsgemäße Optimierungsverfahren anderen Verfahren, die sich artifiziellen Algorithmen zur mRNA-Sekundärstrukturoptimierung bedienen, in der Praxis deutlich überlegen. Es hat sich gezeigt, dass sich
30 mithilfe des erfindungsgemäßen Verfahrens selbst die Expressionsrate von Aminosäuresequenzen, die sich auch ohne Optimierung schon mit sehr guter Löslichkeit im heterologen System exprimieren lassen, durch die erfindungsgemäße

Optimierung anhand der Vireng Genome noch einmal deutlich steigern lässt.

Schließlich kann das Verfahren Schritte umfassen, die dazu
5 dienen, innerhalb der Nukleotidsequenz vorhandene und/oder durch
das Ersetzen von Basentriplets in den Änderungspositionen
zufällig erzeugte Nukleotidfolgen und/oder Motive zu reduzieren
oder auszuschließen, welche die Expression in dem Zielorganismus
nachteilig beeinflussen können. In einigen Ausführungsformen des
10 Verfahrens werden solche ungünstigen Nukleotidfolgen und/oder
Motive zumindest teilweise aus der Nukleotidsequenz entfernt.
Nicht-beschränkende Beispiele für ungünstige Nukleotidfolgen
und/oder Motive sind cis-wirkende mRNA-destabilisierende Motive,
RNase-Spleißstellen, Ribosomen-Bindungsstellen, repetitive
15 Elemente und Erkennungssequenzen von Restriktionsenzymen. Des
Weiteren können z. B. der GC-Gehalt und die mRNA-
Sekundärstruktur der transkribierten Nukleotidsequenz
berücksichtigt werden, um die Nukleotidsequenz zusätzlich für
die Expression in dem Zielorganismus weiter zu verbessern. Wie
20 bereits oben ausgeführt wurde, besteht ein besonderer Vorteil
jedoch darin, dass Nukleotidabfolgen, die für die Expression in
dem Zielorganismus schädlich sind, durch das erfindungsgemäße
Verfahren zum Großteil bereits inhärent ausgeschlossen werden,
insbesondere durch die Länge der Codon-n-Tupel mit n größer
25 gleich 3. Auf diese Weise kann in bestimmten Ausführungsformen
ein zusätzlicher Schritt wie z. B. Optimierung der mRNA-
Sekundärstruktur oder des GC-Gehaltes, Entfernung von mRNA-
destabilisierenden Motiven, Ribosomen-Bindungsstellen,
repetitiver Elemente und/oder Erkennungssequenzen von
30 Restriktionsenzymen von dem Verfahren ausgeschlossen werden.

Gegenstand der vorliegenden Erfindung ist weiterhin eine
Verwendung einer gemäß dem vorbeschriebenen Verfahren

optimierten Nukleotidsequenz zur Herstellung einer synthetischen DNA und/oder zur Proteinexpression in einem Zielorganismus.

Dementsprechend ist ein weiterer Gegenstand der vorliegenden
5 Erfindung ein, insbesondere isoliertes, Nukleinsäuremolekül, das
eine optimierte Nukleotidsequenz umfasst, die durch eines der
hier beschriebenen Verfahren erhalten wurde. Vorzugsweise
handelt es sich bei der Nukleinsäure um DNA. In weiteren
Ausführungsformen wird ein Vektor bereitgestellt, der das,
10 insbesondere isolierte, Nukleinsäuremolekül umfasst.
Erfindungsgemäß optimierte Nukleinsäuremoleküle lassen sich
anhand eines Sequenzvergleichs deutlich von herkömmlich
optimierten Sequenzen unterscheiden. Diesbezüglich wird auch auf
die nachfolgenden Vergleichsbeispiele verwiesen.

15

Ein weiterer Gegenstand der Erfindung ist eine rekombinante
Wirtszelle, die das oben erwähnte, insbesondere isolierte,
Nukleinsäuremolekül oder den oben erwähnten Vektor enthält.

20 Dementsprechend betrifft die vorliegende Erfindung auch ein
Verfahren zur Expression eines, insbesondere rekombinanten,
Proteins in einem Zielorganismus, welches die Bereitstellung
einer gemäß dem obigen Verfahren optimierten Nukleotidsequenz,
welche das Protein codiert, umfasst. Weiterhin kann das
25 Verfahren einen oder mehrere der folgenden Schritte umfassen:
Synthetisieren eines Nukleinsäuremoleküls, welches die
optimierte Nukleinsäuresequenz umfasst; Einbringen des
Nukleinsäuremoleküls in den Zielorganismus; und Kultivieren des
Zielorganismus unter Bedingungen, welche die Expression des
30 Proteins aus der optimierten Nukleinsäuresequenz ermöglichen.
Vorzugsweise wird die Expression zumindest teilweise bei einer
Temperatur kleiner gleich 30 °C, kleiner gleich 25 °C oder
kleiner gleich 20 °C durchgeführt. Es hat sich gezeigt, dass

erfindungsgemäß optimierte Nukleotidsequenzen eine heterologe Proteinexpression bei relativ niedrigen Temperaturen gegenüber herkömmlich optimierten Nukleotidsequenzen deutlich begünstigen. Diesbezüglich wird auch auf die nachfolgenden

5 Ausführungsbeispiele verwiesen. Auf diese Weise eignet sich das erfindungsgemäße Verfahren besonders für die Expression empfindlicher Hochwertproteine und führt gleichzeitig durch Energieeinsparungspotenzial zu mehr Nachhaltigkeit.

10 Ein weiterer Gegenstand der vorliegenden Erfindung ist ein Computerprogramm mit Programmcodemitteln. Die Programmcodemittel des Computerprogramms sind hierbei zur Durchführung eines Verfahrens gemäß vorstehender Beschreibung eingerichtet, wenn das Computerprogramm auf einem Rechner ausgeführt wird. Das
15 Computerprogramm kann eine Schnittstelle zu einer DNA- und/oder RNA-Synthesevorrichtung umfassen.

Gegenstand der vorliegenden Erfindung ist ebenfalls ein computerlesbares Speichermedium, auf dem das vorgenannte
20 Computerprogramm in computerlesbarer Form gespeichert ist.

Ein weiterer Gegenstand der Erfindung ist schließlich eine Vorrichtung zur Optimierung und/oder Herstellung einer Nukleotidsequenz für die Expression einer vorgegebenen
25 Aminosäuresequenz in mindestens einem Zielorganismus. Hierbei weist die Vorrichtung eine Recheneinrichtung auf, die zur Durchführung eines der oben genannten Verfahren eingerichtet ist. Die Vorrichtung kann insbesondere ein DNA- und/oder RNA-Synthesegerät, auch als „DNA/RNA-Synthesizer“ bezeichnet, sein.

30

Im Übrigen versteht es sich, dass sich die bevorzugten und vorteilhaften Ausführungsformen des erfindungsgemäßen Verfahrens, soweit anwendbar, auch auf die anderen Gegenstände

der vorliegenden Erfindung beziehen können. Merkmale, die vorstehend und im Folgenden im Zusammenhang mit dem erfindungsgemäßen Verfahren offenbart sind, können sich daher auch auf die weiteren Gegenstände der Erfindung beziehen und
5 umgekehrt. Die Verwendung des Begriffs „bzw.“ steht im Zweifel für eine „und/oder“-Beziehung.

Kurze Beschreibung der Sequenzen

- 10 SEQ ID NO:1 Nukleotidsequenz der *I. sakaiensis* PETase
(Wildtypsequenz) codierend für die Aminosäuren
(AS) 28-290;
- SEQ ID NO:2 Synthetisch hergestellte Nukleotidsequenz der *I.*
15 *sakaiensis* PETase (codierend für AS 28-290) mit
doppeltem Strep-Tag am C-Terminus nach
herkömmlicher Optimierung für die Expression in *E.*
coli gemäß Stand der Technik (Referenz);
- 20 SEQ ID NO:3 Synthetisch hergestellte Nukleotidsequenz der *I.*
sakaiensis PETase (codierend für AS 28-290) mit
doppeltem Strep-Tag am C-Terminus nach
erfindungsgemäßer Optimierung der Sequenz für die
Expression in *E. coli* mit $n = 2$;
- 25 SEQ ID NO:4 Synthetisch hergestellte Nukleotidsequenz der *I.*
sakaiensis PETase (codierend für AS 28-290) mit
doppeltem Strep-Tag am C-Terminus nach
erfindungsgemäßer Optimierung der Sequenz für die
30 Expression in *E. coli* mit $n = 3$;
- SEQ ID NO:5 Nukleotidsequenz der *A. thaliana* OTP86-DYW Domäne
(AS 826-960) (Wildtypsequenz);

- SEQ ID NO:6 Synthetisch hergestellte Nukleotidsequenz der *A. thaliana* OTP86-DYW Domäne (AS 826-960) mit doppeltem Strep-Tag und *Tobacco etch virus* (TEV)-Schnittstelle am N-Terminus nach herkömmlicher Optimierung für die Expression in *E. coli* gemäß Stand der Technik (Referenz);
- 5
- SEQ ID NO:7 Synthetisch hergestellte Nukleotidsequenz der *A. thaliana* OTP86-DYW Domäne (AS 826-960) mit doppeltem Strep-Tag und *Tobacco etch virus* (TEV)-Schnittstelle am N-Terminus nach erfindungsgemäßer Optimierung für die Expression in *E. coli* mit n = 3;
- 10
- 15
- SEQ ID NO:8 Synthetisch hergestellte Nukleotidsequenz der *A. thaliana* OTP86-DYW Domäne (AS 826-960) mit doppeltem Strep-Tag und *Tobacco etch virus* (TEV)-Schnittstelle am N-Terminus nach erfindungsgemäßer Optimierung für die Expression in *E. coli* anhand von 226 Phagengenomen mit n = 3;
- 20
- SEQ ID NO:9 Proteinsequenz von Citrin (codierend für AS 1-239) als vorgegebene Aminosäuresequenz für die Expression in *H. sapiens*;
- 25
- SEQ ID NO:10 Synthetisch hergestellte Nukleotidsequenz von Citrin (codierend für AS 1-239) mit FLAG-Tag und doppeltem Strep-Tag am N-Terminus nach herkömmlicher Optimierung für die Expression in *H. sapiens* gemäß Stand der Technik (Referenz), 5'-flankiert von einer Kozak-Sequenz (GCCACC);
- 30

- SEQ ID NO:11 Synthetisch hergestellte Nukleotidsequenz von Citrin (codierend für AS 1-239) mit FLAG-Tag und doppeltem Strep-Tag am N-Terminus nach erfindungsgemäßer Optimierung der Sequenz für die Expression in *H. sapiens* mit n = 3, 5'-flankiert von einer Kozak-Sequenz (GCCACC);
- 5
- SEQ ID NO:12 Nukleotidsequenz (Wildtyp) des *H. sapiens* STING1 ER exit protein 1 („STEEP1“) codierend für die AS 1-222;
- 10
- SEQ ID NO:13 Synthetisch hergestellte Nukleotidsequenz von *H. sapiens* STEEP1 (codierend für AS 1-222) mit FLAG-Tag und doppeltem Strep-Tag am N-Terminus nach herkömmlicher Optimierung für die Expression in *H. sapiens* gemäß Stand der Technik (Referenz), 5'-flankiert von einer Kozak-Sequenz (GCCACC);
- 15
- SEQ ID NO:14 Synthetisch hergestellte Nukleotidsequenz von *H. sapiens* STEEP1 (codierend für AS 1-222) mit FLAG-Tag und doppeltem Strep-Tag am N-Terminus nach erfindungsgemäßer Optimierung der Sequenz für die Expression in *H. sapiens* mit n = 3, 5'-flankiert von einer Kozak-Sequenz (GCCACC);
- 20
- 25
- SEQ ID NO:15 Nukleotidsequenz (Wildtyp) des *H. sapiens* nitric oxide synthase-interacting protein (NOSIP) codierend für die AS 1-304;
- 30
- SEQ ID NO:16 Synthetisch hergestellte Nukleotidsequenz von *H. sapiens* NOSIP (codierend für AS 1-304) mit FLAG-Tag und doppeltem Strep-Tag am N-Terminus nach herkömmlicher Optimierung mittels Codon Adaptation

Index und mRNA-Sekundärstrukturoptimierung für die Expression in *H. sapiens* gemäß Stand der Technik (Referenz), 5'-flankiert von einer Kozak-Sequenz (GCCACC);

5

SEQ ID NO:17 Synthetisch hergestellte Nukleotidsequenz von *H. sapiens* NOSIP (codierend für AS 1-304) mit FLAG-Tag und doppeltem Strep-Tag am N-Terminus nach herkömmlicher Optimierung gemäß WO 2020/024917 A1 für die Expression in *Homo sapiens* gemäß Stand der Technik (Referenz), 5'-flankiert von einer Kozak-Sequenz (GCCACC);

10

SEQ ID NO:18 Synthetisch hergestellte Nukleotidsequenz von *H. sapiens* NOSIP (codierend für AS 1-304) mit FLAG-Tag und doppeltem Strep-Tag am N-Terminus nach erfindungsgemäßer Optimierung der Sequenz für die Expression in *H. sapiens* mit n = 3, 5'-flankiert von einer Kozak-Sequenz (GCCACC);

15

20

SEQ ID NO:19 Proteinsequenz von EqFP611 (AS 1-231) als vorgegebene Aminosäuresequenz für die Expression in *S. elongatus*;

25 SEQ ID NO:20 Synthetisch hergestellte Nukleotidsequenz von EqFP611 (codierend für AS 1-231) mit doppeltem Strep-Tag am N-terminus nach herkömmlicher Optimierung für die Expression in *S. elongatus* gemäß Stand der Technik (Referenz), 5'-flankiert von einer Restriktionsschnittstelle *NdeI* und 3'-flankiert von einem Transkriptionsterminator sowie einer *KpnI* Restriktionsschnittstelle;

30

SEQ ID NO:21 Synthetisch hergestellte Nukleotidsequenz von
EqFP611 (codierend für AS 1-231) mit doppeltem
Strep-Tag am N-terminus nach erfindungsgemäßer
Optimierung für die Expression in *S. elongatus* mit
5 n=3, 5'-flankiert von einer
Restriktionsschnittstelle *NdeI* und 3'-flankiert
von einem Transkriptionsterminator sowie einer
KpnI Restriktionsschnittstelle.

10 Kurze Beschreibung der Figuren

Die Erfindung wird im Folgenden anhand von Ausführungsbeispielen
unter Bezugnahme auf die beigefügten Figuren näher erläutert.
Keinesfalls soll die Erfindung auf die Ausführungsbeispiele
15 beschränkt sein. Es zeigen:

Fig. 1 ein Flussdiagramm mit einem schematischen Ablauf
einer Ausführungsform des erfindungsgemäßen
Verfahrens;

20 Fig. 2 SDS-PAGE (A) und quantitative Auswertung (B) der
heterologen Expression von *Ideonella sakaiensis*
PETase in *Escherichia coli* bei 20 °C;

25 Fig. 3 SDS-PAGE (A) und quantitative Auswertung (B) der
heterologen Expression von *Ideonella sakaiensis*
PETase in *Escherichia coli* bei 30 °C;

30 Fig. 4 eine grafische Darstellung der relativen
Häufigkeit der ersten und zweiten Codon-3-Tupel in
einem Ausschnitt der Wildtypsequenz (A) bzw. der
erfindungsgemäß optimierten Nukleotidsequenz (B)
von *Arabidopsis thaliana* OTP86-DYW;

- Fig. 5 SDS-PAGE (A) und quantitative Auswertung (B) der heterologen Expression von *Arabidopsis thaliana* OTP86-DYW in *Escherichia coli* bei 17 °C;
- 5
- Fig. 6 Western-Blot (A) und quantitative Auswertung des Western Blots (B) sowie der Fluoreszenz (C) der heterologen Expression von Citrin in HeLa Zellen;
- 10 Fig. 7 Western-Blot (A) und quantitative Auswertung des Western Blots (B) sowie der Fluoreszenz (C) der heterologen Expression von Citrin in HEK293 Zellen;
- 15 Fig. 8 Western-Blot (A) und quantitative Auswertung (B) der Expression von *H. sapiens* STEEP1 in HeLa Zellen;
- Fig. 9 Western-Blot (A) und quantitative Auswertung (B) der Expression von *H. sapiens* STEEP1 in HEK293 Zellen;
- 20
- Fig. 10 Western-Blot (A) und quantitative Auswertung (B) der Expression von *H. sapiens* NOSIP in HeLa Zellen mit SEQ ID NO:16 (Referenz) und SEQ ID NO:18;
- 25
- Fig. 11 Western-Blot (A) und quantitative Auswertung (B) der Expression von *H. sapiens* NOSIP in HeLa Zellen mit SEQ ID NO:17 (Referenz) und SEQ ID NO:18;
- 30
- Fig. 12 Western-Blot (A) und quantitative Auswertung (B) der Expression von *H. sapiens* NOSIP in HEK293

Zellen mit SEQ ID NO:16 (Referenz) und SEQ ID NO:18.

Detaillierte Beschreibung von Ausführungsbeispielen

5

Die Erfindung wird im Folgenden anhand von Ausführungsbeispielen unter Bezugnahme auf die beigefügten Figuren näher erläutert. Keinesfalls soll die Erfindung auf die Ausführungsbeispiele beschränkt sein.

10

Vergleichsbeispiel 1: Heterologe Expression von *Ideonella sakaiensis* PET-Hydrolase (PETase) in *Escherichia coli*

In einem ersten Vergleichsversuch wurde die Nukleotidsequenz der PETase von *I. sakaiensis*, die für die Aminosäurepositionen 28-290 (Molekulargewicht 27,9 kDa) codiert (SEQ ID NO:1), für die heterologe Expression in *E. coli* nach dem erfindungsgemäßen Verfahren und nach dem Verfahren gemäß WO 2020/024917 A1 als Referenz optimiert.

20

Für die Expression der PETase in *E. coli* als Zielorganismus wurden drei pET28a Expressionsplasmide von der Firma Genscript bezogen, welche die Aminosäuresequenz der PETase mit einem doppelten Strep-Tag am C-Terminus unter einem induzierbaren T7-Promotor codieren. N-terminal wurde die Aminosäuresequenz mit Met (Startcodon) und den Aminosäuren Ala und Ser ergänzt. Jeweils eines der Plasmide enthielt die PETase-codierende Nukleotidsequenz mit doppeltem Strep-Tag nach Optimierung gemäß dem Verfahren aus WO 2020/024917 A1 (SEQ ID NO:2) durch den Optimierungsservice des Herstellers, nach Optimierung gemäß dem erfindungsgemäßen Verfahren mit n = 2 (SEQ ID NO:3) und nach Optimierung gemäß dem erfindungsgemäßen Verfahren mit n = 3 (SEQ ID NO:4), wobei für die erfindungsgemäße Bestimmung der

30

relativen Codon-n-Tupel-Häufigkeit die für Proteine codierenden Teile des *E. coli* Genoms zugrunde gelegt wurden.

Fig. 1 zeigt in diesem Zusammenhang einen schematischen Ablauf
5 einer beispielhaften Durchführung des erfindungsgemäßen
Verfahrens 100 in einer computerimplementierten Ausführungsform
zur Optimierung der PETase aus *I. sakaiensis*. Das Feld 102
repräsentiert hierbei die Eingabe der zu optimierenden
Nukleotidsequenz in den Computer. Selbstverständlich ist es auch
10 möglich, eine Aminosäuresequenz einzugeben, die anschließend in
eine zu optimierende Nukleotidsequenz übersetzt wird. Hierbei
wurde im vorliegenden Beispiel die gesamte codierende Sequenz
durchgehend in Änderungspositionen, d. h. mit je einem Codon
Versatz zwischen benachbarten Änderungspositionen, unterteilt.
15 Auf diese Weise resultierten bezogen auf die Gesamtzahl N der
Aminosäuren der vorgegebenen Aminosäuresequenz N-2 Codon-3-Tupel
bzw. N-1 Codon-2-Tupel. Die Änderungspositionen sind in diesem
Ausführungsbeispiel somit mit der Eingabe der zu optimierenden
Nukleotidsequenz festgelegt.

20

In Feld 104 wurde mithilfe einer DNA-Sequenzdatenbank eine Liste
der proteincodierenden Gene von *E. coli* erstellt. Eine geeignete
DNA-Sequenzdatenbank ist z. B. GenBank (Nucleic Acids Research
41, 2013, D36-42). Eine geeignete Grundlage für die Verknüpfung
25 von Sequenzdaten mit genetischen und funktionellen Informationen
bildet die Reference Sequence (RefSeq) Datenbank (The NCBI
Handbook, 2nd edition, Chapter 18: The Reference Sequence
(RefSeq) Database, Bethesda (MD), National Center for
Biotechnology Information, USA, 2013). Auf dieser Grundlage
30 wurde anschließend die absolute Häufigkeit jedes kombinatorisch
möglichen Codon-n-Tupels sowie die absolute Häufigkeit jedes
kombinatorisch möglichen Aminosäure-n-Tupels innerhalb der

proteincodierenden Gene ermittelt, wobei n in einer Ausführungsform $n = 2$ und in einer Ausführungsform $n = 3$ war.

Mithilfe dieser Informationen wurde in Feld 106 für jedes
5 kombinatorisch mögliche Codon- n -Tupel der Quotient aus seiner absoluten Häufigkeit und der absoluten Häufigkeit des entsprechenden Aminosäure- n -Tupels, für das es codiert, bestimmt und in das weitere Verfahren einbezogen.

- 10 Als optionaler Schritt wurden in Feld 108 unerwünschte Nukleotidsequenzen, wie zum Beispiel die TATA-box „TATAA“ oder die ribosomale Bindestelle „AGGAGG“ von denen dem Fachmann bekannt ist, dass sie die Expression in *E. coli* beeinträchtigen können, eingegeben. Weitere unerwünschte Sequenzmotive waren
15 AAAAAA, TTTTT, AGGAGGT, TATAAA, ATCTGTT, GGAGGT und GGTGGT.

Mithilfe eines computergestützten Optimierungsverfahrens unter Verwendung eines „Simulated Annealing“-Algorithmus wurden anschließend in Feld 110 die Basentriplets in den

- 20 Änderungspositionen der Wildtypsequenz in einer Vielzahl von Interaktionsschritten 112 sukzessive ersetzt, bis die relative Codon- n -Tupel Häufigkeit sämtlicher Codon- n -Tupel in der Nukleotidsequenz einen größtmöglichen gewichteten Mittelwert bei gleichzeitiger Minimierung der Anzahl unerwünschter
25 Nukleotidsequenzen erreichte. Lediglich das Startcodon wurde von der Optimierung ausgeschlossen, wobei es aber grundsätzlich möglich ist, auch das Startcodon und/oder Stopcodon in die Optimierung mit einzubeziehen.

- 30 Die Gewichtung W eines Codon- n -Tupels mit der relativen Häufigkeit P erfolgte mit $W = -1/P$ für P größer 0,0001 und $W = -10.000$ für P kleiner gleich 0,0001. Um den gewichteten Mittelwert F_w zu erhalten, wurden die W sämtlicher Codon- n -Tupel

addiert und dann durch ihre Anzahl $L-(n-1)$ geteilt, wobei L die Anzahl der in die Optimierung einbezogenen Codons in der Nukleotidsequenz ist (hier $L = 864$). Die Wichtung $W = -1/P$ ist degressiv und wurde aufgrund der sehr hohen Werte für sehr
5 kleine P durch -10.000 nach unten begrenzt.

Der gewichtete Mittelwert wurde im vorliegenden Beispiel zusätzlich mit einem Ausdruck für das Vorkommen unerwünschter Sequenzmotive verrechnet. Für jedes unerwünschte Sequenzmotiv
10 wurde ein Wert F_E ermittelt, welcher der Anzahl des unerwünschten Sequenzmotivs in der zu optimierenden Nukleotidsequenz multipliziert mit -1 entspricht. Zusätzlich kann der Wert F_E mit einem individuellen Wert r gewichtet werden, wobei im vorliegenden Beispiel $r = 0,035$ für die Ausführungsform mit $n =$
15 2 und $r = 0,058$ für die Ausführungsform mit $n = 3$ gewählt wurde.

Zur Optimierung der Nukleotidsequenz wurde somit mithilfe des „Simulated Annealing“-Algorithmus nach dem Höchstwert des gewichteten Mittelwerts anhand der Funktion $F = F_W + r \cdot F_E$
20 gesucht, sodass F_W durch die Optimierung einen größtmöglichen Wert annimmt und gleichzeitig die geringstmögliche Anzahl von unerwünschten Sequenzen am Ende der Optimierung in der Nukleotidsequenz enthalten ist.

25 Nach der Optimierung betrug der gewichtete Mittelwert in der erfindungsgemäßen Ausführungsform mit $n = 2$ (SEQ ID NO:3) $F_W = -5,1$. Eine analoge Vergleichsberechnung ergab einen gewichteten Mittelwert der relativen Codon-2-Tupel-Häufigkeiten im Wildtyp (SEQ ID NO:1) $F_W = -13,2$ und für die Referenzsequenz
gemäß WO 2020/024917 A1 (SEQ ID NO:2) $F_W = -7,3$. Der gewichtete
30 Mittelwert in der Ausführungsform mit $n = 3$ betrug nach der erfindungsgemäßen Optimierung (SEQ ID NO:4) $F_W = -10,4$. Eine analoge Vergleichsberechnung ergab einen gewichteten Mittelwert

der relativen Codon-3-Tupel-Häufigkeiten im Wildtyp (SEQ ID NO:1) $F_W = -526,8$ und in der Referenzsequenz gemäß WO 2020/024917 A1 (SEQ ID NO:2) $F_W = -336,4$.

- 5 Nachdem das Zielkriterium erreicht wurde, erfolgte in Feld 114 die Ausgabe der optimierten Nukleotidsequenzen SEQ ID NO:3 und SEQ ID NO:4, die anschließend entsprechend synthetisiert wurden.

Es ist aus dem Sequenzprotoll unmittelbar zu erkennen, dass sich
10 die erfindungsgemäß optimierte Nukleotidsequenz von der Wildtypsequenz bereits auf Nukleotidebene deutlich unterscheidet. So hat die Wildtypsequenz (SEQ ID NO:1) nur 82,3% identische Nukleotide mit der gemäß dem erfindungsgemäßen Verfahren in der Ausführungsform mit $n = 2$ optimierten Sequenz
15 (SEQ ID NO:3) bzw. nur 82,9% identische Nukleotide mit der optimierten Sequenz aus dem erfindungsgemäßen Verfahren in der Ausführungsform mit $n = 3$ (SEQ ID NO:4). Ein ebenfalls sehr deutlicher Unterschied besteht darüber hinaus auch zwischen den nach dem erfindungsgemäßen Verfahren optimierten
20 Nukleotidsequenzen und der nach dem Stand der Technik optimierten Sequenz. Die Sequenz, die gemäß WO 2020/024917 A1 als Referenz optimiert wurde (SEQ ID NO:2), besitzt nur 89,2% identische Nukleotide mit der optimierten Sequenz aus dem erfindungsgemäßen Verfahren in der Ausführungsform mit $n = 2$
25 (SEQ ID NO:3) und nur 84,7% identische Nukleotide mit der optimierten Sequenz aus dem erfindungsgemäßen Verfahren in der Ausführungsform mit $n = 3$ (SEQ ID NO:4). Diese großen Unterschiede sind angesichts der vorgegebenen Aminosäuresequenz aus fachmännischer Sicht überraschend und zeigen, dass das
30 erfindungsgemäße Optimierungsverfahren zu einem völlig anderen Ergebnis führt als das Verfahren gemäß dem aktuellen Stand der Technik.

Als Expressionssystem wurde der *E. coli* Stamm BL21 verwendet (New England Biolabs GmbH, Frankfurt am Main, Deutschland). Dazu wurden jeweils drei verschiedene Aliquote des Expressionswirtes mit einem der Expressionsplasmide durch Elektroporation mit 0,1
5 µg DNA transformiert und auf eine LB-Agar Platte mit 100 mg/mL Kanamycin ausgestrichen. Die Platten wurden über Nacht bei 37 °C inkubiert. Aus mehreren Kolonien von jeweils einer Platte wurden Vorkulturen in 50 mL LB-Medium mit 100 mg/mL Kanamycin angesetzt und über Nacht bei 37 °C im Schüttler bei 180 U/min inkubiert.

10

Die Expressionskulturen wurden aus jeweils 1 mL Vorkultur und 99 mL TB Medium angesetzt. Die Expression der rekombinanten PETase in den Kulturen wurden bei einer OD600 von 0,6 durch Zugabe von IPTG in einer Endkonzentration von 1 mM induziert. Die
15 Expression erfolgte in einer Variante bei 20 °C für 14 Stunden und in einer weiteren Variante bei 30 °C für fünf Stunden.

20

Vor der Zellernte wurde die OD600 der Expressionskulturen bestimmt und jeweils die gleiche Zellmenge von jeder der
Kulturen geerntet, um die Proteinausbeuten anhand der Zellmasse zu normalisieren. Die Zellpellets wurden in 10 mL Puffer A (20 mM Tris-Cl, pH 7.5, 150 mM NaCl, 1 mM DTT) gelöst und mithilfe von Ultraschall aufgeschlossen. Das Zelllysate wurde anschließend bei 20.000 g für eine Stunde zentrifugiert, um die unlöslichen
25 Zellbestandteile als Pellet abzutrennen.

30

Der Überstand mit der löslichen Fraktion wurde in einem Eppendorf-Gefäß mit 200 µL in Puffer A äquilibrierten Streptactin-Beads (IBA Lifesciences, Göttingen, Deutschland) vermischt. Die Beads wurden im Eppendorfgefäß durch Zentrifugation und Entfernen des Überstandes zweimal mit 1 mL Puffer A gewaschen. Die gebundenen Proteine wurden mit 200 µL Puffer A mit 10 mM Desthiobiotin eluiert.

Die Identität des Proteins wurde durch SDS-Polyacrylamidgelelektrophorese (SDS-PAGE) analytisch verifiziert. Die Proteinmenge in den SDS-PAGE-Gelbanden wurde mithilfe der ImageJ-Software (National Institutes of Health, USA) quantifiziert. Zusätzlich erfolgte eine Bestimmung der Proteinkonzentration in den jeweiligen Überständen mithilfe des Bradford Assays (Thermo Fisher Scientific, Bremen, Deutschland) nach Herstellerangaben durchgeführt.

Die Ergebnisse sind in Fig. 2 und Fig. 3 gezeigt. Fig. 2A zeigt eine Abbildung der SDS-PAGE-Analyse von der Expression bei 20 °C. Bahn 1 enthält einen Größenmarker, Bahn 2 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:2 (Referenz), Bahn 3 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:3 und Bahn 4 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:4. Die PETase wurde von allen drei Plasmiden ausweislich der Bande bei 30,4 kDa, die dem Molekulargewicht des Proteins einschließlich des doppelten Strep-Tags entspricht, erfolgreich exprimiert. Es ist jedoch bereits anhand der Bandenstärke deutlich erkennbar, dass das Plasmid mit der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:3, wobei für die Optimierung $n = 2$ gewählt wurde, zu einer signifikant gesteigerten Proteinausbeute im Vergleich zum Plasmid mit der herkömmlich optimierten Nukleotidsequenz SEQ ID NO:2 führte. Das Plasmid mit der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:4, wobei für die Optimierung $n = 3$ gewählt wurde, führte insgesamt zur höchsten Ausbeute an rekombinanter PETase.

In dem in Fig. 2B gezeigten Säulendiagramm ist die relative Proteinausbeute an löslicher PETase in Abhängigkeit von der jeweils verwendeten Nukleotidsequenz dargestellt, wobei die

quantitativ ermittelte Proteinmenge jeweils auf die Proteinmenge aus dem Referenzversuch mit SEQ ID NO:2 normiert wurde. Die schraffierten Säulen zeigen das Ergebnis der Quantifizierung anhand der SDS-PAGE, die weißen Säulen zeigen das Ergebnis der Quantifizierung durch den Bradford-Assay. Die Quantifizierung ergab, dass die erfindungsgemäße Optimierung der PETase-Nukleotidsequenz mit $n = 2$ die Ausbeute an löslichem Protein gegenüber der Referenzoptimierung verdoppelte, während die erfindungsgemäße Optimierung mit $n = 3$ die Proteinausbeute sogar etwa versechsfachte.

Fig. 3A zeigt eine Abbildung der SDS-PAGE-Analyse von der Expression bei 30 °C. Bahn 1 enthält einen Größenmarker, Bahn 2 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:2 (Referenz) und Bahn 3 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:4. Die PETase wurde auch hier von beiden Plasmiden ausweislich der Bande bei 30,4 kDa erfolgreich exprimiert. Anhand der Bandenstärke ist zu erkennen, dass das Plasmid mit der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:4 zu einer höheren Proteinausbeute führte als das Plasmid mit der Referenzsequenz SEQ ID NO:2.

In Fig. 3B ist wiederum die relative Proteinausbeute an löslicher PETase in Abhängigkeit von der jeweils verwendeten Nukleotidsequenz dargestellt, wobei die quantitativ ermittelte Proteinmenge auf die Proteinmenge aus dem Referenzversuch mit SEQ ID NO:2 normiert wurde. Die schraffierten Säulen zeigen das Ergebnis der Quantifizierung anhand der SDS-PAGE, die weißen Säulen zeigen das Ergebnis der Quantifizierung durch den Bradford-Assay. Die quantitative Analyse belegt eine mehr als verdoppelte Expression der PETase mithilfe der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:4 im Vergleich zur gemäß WO 2020/024917 A1 optimierten Sequenz.

Eine derart signifikante Verbesserung der heterologen Expressionsrate eines Proteins mithilfe des erfindungsgemäßen Verfahrens zur Optimierung der proteincodierenden Nukleotidsequenz im Vergleich zum aktuellen Stand der Technik ist für den Fachmann nicht zu erwarten gewesen. Hierbei überrascht, dass bereits der niedrigste vorgesehene Parameter $n = 2$ eine fast 200%ige Steigerung der Expressionsrate gegenüber dem Stand der Technik bewirkte und damit die Überlegenheit des erfindungsgemäßen Optimierungsverfahrens gegenüber den etablierten Konzepten nachdrücklich belegt.

Vergleichsbeispiel 2: Heterologe Expression von *Arabidopsis thaliana* OTP86-DYW in *Escherichia coli*

In einem weiteren Vergleichsversuch wurde die Nukleotidsequenz der OTP86-DYW Domäne in den Aminosäurepositionen 826-960 von *A. thaliana* (SEQ ID NO:5) für die heterologe Expression in *E. coli* nach dem erfindungsgemäßen Verfahren und nach dem Verfahren gemäß WO 2020/024917 A1 als Referenz optimiert. Es handelt sich bei der OTP86-DYW Domäne um ein empfindliches Pflanzenprotein, das bekanntermaßen schwierig im heterologen System zu exprimieren ist.

Für die Expression der OTP86-DYW Domäne in *E. coli* als Zielorganismus wurden drei pET41 Expressionsplasmide kloniert, welche die Aminosäuresequenz 826-960 der OTP86-DYW Domäne mit einem doppelten Strep-Tag und einer TEV-Protease Spaltstelle als Insert unter einem induzierbaren T7-Promotor enthielten. Am N-Terminus der Aminosäuresequenz wurde außerdem ein Met (Startcodon) und ein Gly ergänzt. Die Inserts wurden von der Firma Genscript bezogen. Jeweils eines der Plasmide enthielt die OTP86-DYW-codierende Nukleotidsequenz nach Optimierung gemäß dem

Verfahren aus WO 2020/024917 A1 (SEQ ID NO:6) durch den Optimierungsservice des Herstellers, nach Optimierung gemäß dem erfindungsgemäßen Verfahren mit dem Parameter $n = 3$ (SEQ ID NO:7), wobei für die erfindungsgemäße Bestimmung der relativen Codon-3-Tupel-Häufigkeit der codierende Teil des *E. coli* Genoms zugrunde gelegt wurde, und nach Optimierung gemäß dem erfindungsgemäßen Verfahren mit $n = 3$ (SEQ ID NO:8), wobei für die erfindungsgemäße Bestimmung der relativen Codon-3-Tupel-Häufigkeit der codierende Teil der Genome der folgenden zur Infektion von *E. coli* befähigten Viren bzw. Phagen zugrunde gelegt wurde:

Enterobacteria phage 13a, Enterobacteria phage 285P, Enterobacteria phage 933W, Enterobacteria phage 9g, Enterobacteria phage BA14, Enterobacteria phage BP-4795, Enterobacteria phage Bp7, Enterobacteria phage EcoDS1, Enterobacteria phage G4, Enterobacteria phage GA, Enterobacteria phage GEC-3S, Enterobacteria phage HK106, Enterobacteria phage HK140, Enterobacteria phage HK225, Enterobacteria phage ID2 Moscow/ID/2001, Enterobacteria phage IME08, Enterobacteria phage IME10, Enterobacteria phage If1, Enterobacteria phage Ike, Enterobacteria phage J8-65, Enterobacteria phage JS10, Enterobacteria phage JenK1, Enterobacteria phage JenP1, Enterobacteria phage JenP2, Enterobacteria phage K1F, Enterobacteria phage M, Enterobacteria phage MS2, Enterobacteria phage MX1, Enterobacteria phage P4, Enterobacteria phage P88, Enterobacteria phage PRD1, Enterobacteria phage Phil, Enterobacteria phage RB27, Enterobacteria phage RB49, Enterobacteria phage RB51, Enterobacteria phage RB68, Enterobacteria phage RB69, Enterobacteria phage SP, Enterobacteria phage ST104, Enterobacteria phage Sf101, Enterobacteria phage SfI, Enterobacteria phage SfV, Enterobacteria phage St-1, Enterobacteria phage T3,

Enterobacteria phage T7, Enterobacteria phage UAB_Phi20,
Enterobacteria phage UAB_Phi78, Enterobacteria phage VT2-Sakai,
Enterobacteria phage VT2phi_272, Enterobacteria phage WA13,
Enterobacteria phage YYZ-2008, Enterobacteria phage alpha3,
5 Enterobacteria phage cdtI, Enterobacteria phage fd,
Enterobacteria phage fiAA91-ss, Enterobacteria phage mEp043 c-1,
Enterobacteria phage mEp235, Enterobacteria phage mEp237,
Enterobacteria phage mEp460, Enterobacteria phage phi80,
Enterobacteria phage phi92, Enterobacteria phage phiEcoM-GJ1,
10 Enterobacteria phage phiP27, Enterobacteria phage vB_EcoM_VR5,
Enterobacteria phage vB_EcoP_ACG-C91, Enterobacteria phage
vB_EcoS_NBD2, Enterobacteria phage vB_EcoS_Roguel,
Enterobacteria phage vB_KleM-RaK2, Escherichia phage 121Q,
Escherichia phage 172-1, Escherichia phage 4MG, Escherichia
15 phage 64795_ec1, Escherichia phage ADB-2, Escherichia phage
APCEc01, Escherichia phage AR1, Escherichia phage Av-05,
Escherichia phage Bp4, Escherichia phage CAjan, Escherichia
phage CICC 80001, Escherichia phage D108, Escherichia phage
EB49, Escherichia phage EC6, Escherichia phage ECBP1,
20 Escherichia phage ECBP2, Escherichia phage ECBP5, Escherichia
phage ECML-117, Escherichia phage ECML-134, Escherichia phage
ECML-4, Escherichia phage EK99P-1, Escherichia phage Envy,
Escherichia phage FFH2, Escherichia phage FV3, Escherichia phage
Gluttony, Escherichia phage HK446, Escherichia phage HK542,
25 Escherichia phage HK544, Escherichia phage HK578, Escherichia
phage HK629, Escherichia phage HK630, Escherichia phage HK633,
Escherichia phage HK639, Escherichia phage HK75, Escherichia
phage HX01, Escherichia phage HY01, Escherichia phage HY02,
Escherichia phage HY03, Escherichia phage IME11, Escherichia
30 phage JES2013, Escherichia phage JH2, Escherichia phage JS98,
Escherichia phage JSE, Escherichia phage K1-dep(1), Escherichia
phage K1-dep(4), Escherichia phage KBNP1711, Escherichia phage
LM33_P1, Escherichia phage Lw1, Escherichia phage MX01,

Escherichia phage Min27, Escherichia phage NJ01, Escherichia
phage P13374, Escherichia phage P483, Escherichia phage P694,
Escherichia phage PA2, Escherichia phage PBECO 4, Escherichia
phage PE3-1, Escherichia phage PhaxI, Escherichia phage Pollock,
5 Escherichia phage QL01, Escherichia phage RB3, Escherichia phage
SUSP1, Escherichia phage SUSP2, Escherichia phage Seurat,
Escherichia phage Stx2 II, Escherichia phage TL-2011b,
Escherichia phage TL-2011c, Escherichia phage Ufv-AREG1,
Escherichia phage V5, Escherichia phage WG01, Escherichia phage
10 YD-2008.s, Escherichia phage e4/1c, Escherichia phage ime09,
Escherichia phage mEp234, Escherichia phage mEpX1, Escherichia
phage mEpX2, Escherichia phage phAPEC8, Escherichia phage
phi191, Escherichia phage phiK, Escherichia phage phiKT,
Escherichia phage phiV10, Escherichia phage pro147, Escherichia
15 phage pro483, Escherichia phage slur01, Escherichia phage
slur02, Escherichia phage slur05, Escherichia phage slur14,
Escherichia phage slur16, Escherichia phage vB_EcoM-Ufv13,
Escherichia phage vB_EcoM-VpaE1, Escherichia phage vB_EcoM-ep3,
Escherichia phage vB_EcoM_112, Escherichia phage vB_EcoM_ACG-
20 C40, Escherichia phage vB_EcoM_AYO145A, Escherichia phage
vB_EcoM_Alf5, Escherichia phage vB_EcoM_ECO1230-10, Escherichia
phage vB_EcoM_JS09, Escherichia phage vB_EcoM_PhAPEC2,
Escherichia phage vB_EcoM_VR20, Escherichia phage vB_EcoM_VR25,
Escherichia phage vB_EcoM_VR26, Escherichia phage vB_EcoM_VR7,
25 Escherichia phage vB_EcoP_24B, Escherichia phage vB_EcoP_G7C,
Escherichia phage vB_EcoP_GA2A, Escherichia phage
vB_EcoP_PhAPEC5, Escherichia phage vB_EcoP_PhAPEC7, Escherichia
phage vB_EcoP_SU10, Escherichia phage vB_EcoS_AHP42, Escherichia
phage vB_EcoS_AHS24, Escherichia phage vB_EcoS_AKS96,
30 Escherichia phage vB_EcoS_FFH1, Escherichia phage vB_Eco_ACG-
M12, Escherichia phage wV7, Escherichia phage wV8, Escherichia
virus 186, Escherichia virus AKFV33, Escherichia virus CBA120,
Escherichia virus DT57C, Escherichia virus EPS7, Escherichia

virus HK022, Escherichia virus HK97, Escherichia virus I22, Escherichia virus JL1, Escherichia virus K1-5, Escherichia virus K1E, Escherichia virus Lambda, Escherichia virus M13, Escherichia virus Mu, Escherichia virus N15, Escherichia virus N4, Escherichia virus P1, Escherichia virus P2, Escherichia virus RB16, Escherichia virus RB32, Escherichia virus Rtp, Escherichia virus SSL2009a, Escherichia virus T1, Escherichia virus T4, Escherichia virus T5, Escherichia virus TLS, Escherichia virus Wphi, Escherichia virus phiEco32, Escherichia virus phiX174. Die Virengenome wurden aus der Refseq-Datenbank am 31. Mai 2019 abgerufen.

Im Übrigen erfolgte die Versuchsdurchführung für die erfindungsgemäße Sequenzoptimierung im Wesentlichen wie in Vergleichsbeispiel 1 beschrieben. Für den Fall der Optimierung auf Basis der Phagengenome, welche kumulativ eine deutlich geringere Größe als das Genom von *E. coli* besitzen, wurden die relativen Codon-3-Tupel-Häufigkeiten P als Mittelpunkte des Clopper-Pearson-Vertrauensintervalls mit einem Konfidenzniveau von 95% berechnet. Weiterhin waren die unerwünschten Sequenzmotive in diesem Beispiel AAAAAA, TTTTT, AGGAGGT, TATAAA, ATCTGTT, GGAGGT, GGTGGT, CCATGG und AAGCTT, deren jeweilige Anzahl für die Optimierung auf Basis des *E. coli* Genoms sowie auf Basis der Phagengenome mit $r = 0,06$ gewichtet wurde. Die Anzahl L der in die Optimierung einbezogenen Codons in der Nukleotidsequenz ist in diesem Beispiel 501, da das Startcodon und das darauffolgende Glycincodon nicht in die Optimierung einbezogen wurden. Die Expression wurde bei 17 °C durchgeführt, wobei der Puffer A kein DTT enthielt.

30

Nach der Optimierung betrug der gewichtete Mittelwert in der erfindungsgemäßen Ausführungsform auf Basis des *E. coli* Genoms mit $n = 3$ (SEQ ID NO:7) $F_w = -7,1$. Eine analoge

Vergleichsberechnung ergab einen gewichteten Mittelwert der relativen Codon-3-Tupel-Häufigkeiten im Wildtyp (SEQ ID NO:5) bezogen auf das *E. coli* Genom $F_w = -2533.5$ und für die Referenzsequenz gemäß WO 2020/024917 A1 (SEQ ID NO:6)

5 $F_w = -451,4$. Der gewichtete Mittelwert in der Ausführungsform auf Basis der Genome der zur Infektion von *E. coli* befähigten Viren bzw. Phagen mit $n = 3$ betrug nach der erfindungsgemäßen Optimierung (SEQ ID NO:8) $F_w = -9,2$. Eine analoge Vergleichsberechnung ergab einen gewichteten Mittelwert der

10 relativen Codon-3-Tupel-Häufigkeiten im Wildtyp (SEQ ID NO:5) auf Basis der Viren- bzw. Phagengenome $F_w = -56,8$ und für die Referenzsequenz gemäß WO 2020/024917 A1 (SEQ ID NO:6) $F_w = -37,2$.

Das Sequenzprotokoll verdeutlicht auch hier, dass bereits auf

15 Nukleotidebene erhebliche Unterschiede zwischen der Nukleotidsequenz vom Wildtyp und den Nukleotidsequenzen, die gemäß dem erfindungsgemäßen Verfahren optimiert wurden, bestehen. So hat die Wildtypsequenz (SEQ ID NO:5) nur 72,7% identische Nukleotide mit der optimierten Sequenz aus dem

20 erfindungsgemäßen Verfahren in der Ausführung mit $n = 3$ (SEQ ID NO:7), wobei für die erfindungsgemäße Bestimmung der relativen Codon-3-Tupel-Häufigkeit der codierende Teil des *E. coli* Genoms zugrunde gelegt wurde, und nur 76,9% identische Nukleotide mit der optimierten Sequenz aus dem erfindungsgemäßen Verfahren in

25 der Ausführung mit $n = 3$ (SEQ ID NO:8), wobei für die erfindungsgemäße Bestimmung der relativen Codon-3-Tupel-Häufigkeit der codierende Teil der Genome der zur Infektion von *E. coli* befähigten Viren bzw. Phagen zugrunde gelegt wurde.

30 Ein sehr deutlicher Unterschied besteht darüber hinaus wiederum auch zwischen den nach dem erfindungsgemäßen Verfahren optimierten Nukleotidsequenzen und der nach dem Stand der Technik optimierten Sequenz. Die Sequenz, die nach dem Stand der

Technik gemäß WO 2020/024917 A1 als Referenz optimiert wurde (SEQ ID NO:6), besitzt nur 85,2% identische Nukleotide mit der optimierten Sequenz gemäß dem erfindungsgemäßen Verfahren anhand des *E. coli* Genoms (SEQ ID NO:7) und nur 74,6% identische Nukleotide mit der optimierten Sequenz gemäß dem erfindungsgemäßen Verfahren anhand der Genome der zur Infektion von *E. coli* befähigten Viren bzw. Phagen (SEQ ID NO:8).

Fig. 4 zeigt die relative Häufigkeit der ersten Codon-3-Tupel in der Wildtypsequenz SEQ ID NO:5 (A) und die relative Häufigkeit der zweiten Codon-3-Tupel in der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:7 (B) von OTP86-DYW für den Sequenzabschnitt, der im Sequenzprotokoll jeweils den Nukleotidpositionen 211-330 entspricht. Der gezeigte Sequenzabschnitt enthält die Codons Nummer 71 bis einschließlich Nummer 110 von OTP86-DYW. Jeweils drei benachbarte Codons bilden eine Änderungsposition mit einem Codon-3-Tupel, wobei der gezeigte Sequenzabschnitt durchgehend in 38 Änderungspositionen bzw. Codon-3-Tupel unterteilt wurde, die hier als n_1 bis n_{38} bezeichnet sind. Zwischen aufeinanderfolgenden Änderungspositionen liegt jeweils ein Versatz von einem Codon. In den Diagrammen ist jedem Codon-3-Tupel (x-Achse) durch einen horizontalen Strich die entsprechende relative Häufigkeit in Prozent (y-Achse) zugeordnet, mit der das jeweilige Codon-3-Tupel das entsprechende Aminosäure-3-Tupel in den proteincodierenden Genen von *E. coli* codiert. Die vertikalen Striche zeigen jeweils die Spanne der relativen Häufigkeiten sämtlicher für eine bestimmte Änderungsposition in Betracht kommender Codon-3-Tupel, die für das entsprechende Aminosäure-n-Tupel der Änderungsposition codieren.

Es ist ersichtlich, dass durch das erfindungsgemäße Verfahren eine deutliche Erhöhung der relativen Codon-3-Tupel-Häufigkeit

in einem Großteil der gezeigten Änderungspositionen stattfand. Im Ergebnis wurde in jeder der gezeigten Änderungspositionen bis auf Position n_{25} jeweils mindestens eines der Basentriplets durch ein synonymes Basentriplett ersetzt; mehrheitlich wurden

5 in den Änderungspositionen jeweils gleich zwei oder drei Basentriplets durch ein synonymes Basentriplett ersetzt, um die relative Häufigkeit der zweiten Codon-3-Tupel optimal zu erhöhen. Es ist ebenfalls erkennbar, dass nur in einem Teil der Änderungspositionen das zweite Codon-3-Tupel jeweils dem Codon-

10 3-Tupel mit der größten relativen Häufigkeit in *E. coli* entspricht. Ferner wurde z. B. in der Änderungsposition n_{35} ein zweites Codon-3-Tupel gebildet, das eine geringere relative Häufigkeit als das ursprüngliche erste Codon-3-Tupel aufweist, um dadurch die relativen Häufigkeiten der kritischeren ersten

15 Codon-3-Tupel in den anderen Änderungspositionen erhöhen zu können. Auf diese Weise konnte ein größtmöglicher gewichteter Mittelwert der relativen Codon-3-Tupel-Häufigkeiten erreicht werden.

20 Die Ergebnisse der zugehörigen heterologen Expression in *E. coli* sind in Fig. 5 gezeigt. Fig. 5A zeigt eine Abbildung der SDS-PAGE-Analyse von der Expression bei 17 °C. Bahn 1 enthält einen Größenmarker, Bahn 2 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:6 (Referenz), Bahn 3 enthält das

25 Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:7 und Bahn 4 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:8. Die OTP86-DYW Domäne wurde von allen drei Plasmiden ausweislich der Bande bei 19,5 kDa erfolgreich exprimiert. Ein Vergleich der Bandenstärken lässt bereits

30 erkennen, dass die Plasmide mit den erfindungsgemäß optimierten Nukleotidsequenzen SEQ ID NO:7 und SEQ ID NO:8 zu deutlich höheren Proteinausbeuten an rekombinanter OTP86-DYW Domäne

führen als das Plasmid mit der herkömmlich optimierten Nukleotidsequenz SEQ ID NO:6.

In dem in Fig. 5B gezeigten Säulendiagramm ist die relative Proteinausbeute an löslicher OTP86-DYW Domäne in Abhängigkeit von der jeweils verwendeten Nukleotidsequenz dargestellt, wobei die quantitativ ermittelte Proteinmenge jeweils auf die Proteinmenge aus dem Referenzversuch mit SEQ ID NO:6 normiert wurde. Die schraffierten Säulen zeigen das Ergebnis der Quantifizierung anhand der SDS-PAGE, die weißen Säulen zeigen das Ergebnis der Quantifizierung durch photometrische UV-Absorptionsmessung bei 260 und 280 nm (Nanodrop). Die Quantifizierung bestätigte, dass die erfindungsgemäße Optimierung der Nukleotidsequenz unter Zugrundelegung des *E. coli* Transkriptoms mit dem Parameter $n = 3$ nahezu die dreifache Ausbeute an löslichem Protein gegenüber der Referenzoptimierung ergab, während die erfindungsgemäße Optimierung unter Zugrundelegung der Viren- bzw. Phagengenome mit dem Parameter $n = 3$ die Proteinausbeute nahezu verdoppelte.

Eine derart signifikante Verbesserung der heterologen Expressionsrate eines Proteins mithilfe des erfindungsgemäßen Verfahrens zur Optimierung der proteincodierenden Nukleotidsequenz im Vergleich zum aktuellen Stand der Technik ist für den Fachmann wiederum nicht zu erwarten gewesen.

Vergleichsbeispiel 3: Heterologe Expression von Citrin in *H. sapiens* (HeLa) Zellkultur

In einem weiteren Vergleichsversuch wurde die Nukleotidsequenz des fluoreszierenden Proteins Citrin (SEQ ID NO:9) für die heterologe Expression in humanen *HeLa* Zellen als Zielorganismus nach dem erfindungsgemäßen Verfahren und nach einem

herkömmlichen, auf dem Codon Adaptation Index und lokaler mRNA Sekundärstrukturoptimierung basierenden Verfahren als Referenz optimiert. Citrin ist eine Variante des grün fluoreszierenden Proteins (GFP) aus *Aequorea victoria* und wird häufig für
5 Reporter-Assays und in der Fluoreszenzmikroskopie verwendet.

Für die Expression von Citrin in HeLa Zellen wurden zwei pTwist CMV Expressionsplasmide von der Firma Twist Bioscience (San Francisco, CA, USA) bezogen, welche die Aminosäuresequenz von
10 Citrin mit einem FLAG-Tag gefolgt von einem doppelten Strep-Tag am N-Terminus unter einem konstitutiven Cytomegalovirus-Promoter codieren. N-terminal wurde die Aminosäuresequenz mit Met (Startcodon) und der Aminosäure Ala ergänzt. Vor dem Startcodon wurde eine Kozak-Sequenz eingefügt. Eines der Plasmide enthielt
15 die Citrin-codierende Nukleotidsequenz mit FLAG- und doppeltem Strep-Tag nach Optimierung durch das Verfahren des Herstellers (SEQ ID NO:10) als Referenz. Das andere Plasmid enthielt die Citrin-codierende Nukleotidsequenz mit FLAG- und doppeltem Strep-Tag nach Optimierung gemäß dem erfindungsgemäßen Verfahren
20 mit $n = 3$ (SEQ ID NO:11), wobei für die erfindungsgemäße Bestimmung der relativen Codon-n-Tupel-Häufigkeit die für Proteine codierenden Teile des *H. sapiens* Genoms zugrunde gelegt wurden.

25 Im Übrigen erfolgte die Versuchsdurchführung für die erfindungsgemäße Sequenzoptimierung im Wesentlichen wie in Vergleichsbeispiel 1 beschrieben. Analog zur Optimierung für *E. coli* als Wirt wurden für *H. sapiens* als optionaler Schritt in Feld 108 die unerwünschten Sequenzmotive wie z. B. die TATA-Box
30 TATAAA, von denen dem Fachmann bekannt ist, dass sie die Expression in *H. sapiens* beeinträchtigen können, eingegeben. Weitere unerwünschte Sequenzmotive waren ATTTA, GCCACC, GCCGCC, AATAAA und ATTAAA, deren jeweilige Anzahl mit $r = 0,62$ gewichtet

wurde. Die Anzahl L der in die Optimierung einbezogenen Codons in der Nukleotidsequenz ist in diesem Beispiel 819, die Kozak Sequenz am 5'-Ende wurde nicht optimiert.

- 5 Es ist aus dem Sequenzprotoll unmittelbar zu erkennen, dass sich die erfindungsgemäß optimierte Nukleotidsequenz (SEQ ID NO:11) von der nach dem Stand der Technik optimierten Sequenz (SEQ ID NO:10) bereits auf Nukleotidebene deutlich unterscheidet. So hat die gemäß dem Stand der Technik optimierte Sequenz (SEQ ID
- 10 NO:10) nur 81,4% identische Nukleotide mit der gemäß dem erfindungsgemäßen Verfahren in der Ausführungsform mit $n = 3$ optimierten Sequenz (SEQ ID NO:11). Diese großen Unterschiede sind angesichts der vorgegebenen Aminosäuresequenz aus fachmännischer Sicht überraschend und zeigen, dass das
- 15 erfindungsgemäße Optimierungsverfahren auch für eine Optimierung in Säugerzellen, hier *Homo sapiens*, zu einem völlig anderen Ergebnis führt als das Verfahren gemäß dem aktuellen Stand der Technik.
- 20 Nach der Optimierung betrug der gewichtete Mittelwert in der erfindungsgemäßen Ausführungsform mit $n = 3$ (SEQ ID NO:11) $F_w = -9,6$. Eine analoge Vergleichsberechnung ergab einen gewichteten Mittelwert der relativen Codon-3-Tupel-Häufigkeiten für die Referenzsequenz (SEQ ID NO:10) von $F_w = -41,8$.
- 25 Für die Expression von Citrin wurden HeLa-Zellen 24 Stunden vor der Transfektion in 6-Well-Platten mit DMEM High glucose Medium (Biowest SAS, Nuaille, Frankreich) mit 10% FCS (Biochrom AG - Berlin, Deutschland) und 1% Penicillin/Streptomycin (Biowest)
- 30 übertragen. Die Transfektionen wurden mit 2 μ g Plasmid und Rotifect (Carl Roth GmbH, Karlsruhe, Deutschland) nach Herstellerangaben durchgeführt. 70 Stunden nach der Transfektion wurde das Medium entfernt, die Zellen wurden mit 1 mL

eisgekühlter phosphatgepufferter Salzlösung (PBS) gewaschen und in RIPA Lyse Puffer resuspendiert. Die Lysate wurden mit 6-fach SDS Ladepuffer versetzt und auf einem 15% SDS Polyacrylamidgel nach Größe aufgetrennt. Die Proteinproben auf dem Gel wurden

5 anschließend mittels Western-Blotting auf eine Nitrozellulosemembran übertragen. Die unspezifischen Bindestellen der Membran wurden mit 2% BSA geblockt und die Membran wurde mit den Primärantikörpern gegen das FLAG-getaggte exprimierte Zielprotein oder das Housekeeping Gen GAPDH

10 (Ladekontrolle) über Nacht inkubiert. Die Membran wurde mit TBS Tween gewaschen und mit dem Meerrettichperoxidase (HRP) gekoppeltem Sekundärantikörper gegen Kaninchen (FLAG) oder Maus (GAPDH) inkubiert. Die Proteine wurden mit dem ECL-Kit (Pierce, Waltham, MA, USA) visualisiert und die Banden mit ImageQuantTL

15 (Cytiva, Marlborough, MA, USA) quantifiziert. Zur vergleichenden Auswertung der Proteinexpression wurde die Bandenstärke von Citrin in das jeweilige Verhältnis zur Bandenstärke der Ladekontrolle GAPDH gesetzt, um auf diese Weise die aufgetragene Proteinmenge in Bezug auf die Zellmenge zu normalisieren.

20

Weiterhin wurden die Zellysate für zwei Minuten bei 13.000 g zentrifugiert und die Fluoreszenz von Citrin im Überstand in einem Tecan Spark Plate Reader bei einer Anregungswellenlänge von 516 nm und einer Emissionswellenlänge von 529 nm in

25 Triplikaten gemessen. Die Intensität der Fluoreszenzen wurde wiederum in Relation zu der jeweiligen Bandenintensität der Ladekontrollen (GAPDH) gesetzt, um den unterschiedlichen Zelldichten der Kulturen Rechnung zu tragen.

30

Die Ergebnisse sind in Fig. 6 gezeigt. Fig. 6A zeigt eine Abbildung des Western Blots, gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für das FLAG getaggte Citrin sowie gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für die Ladekontrolle

GAPDH. Bahn 1 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:10 (Referenz), Bahn 2 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:11, die erfindungsgemäß optimiert wurde. Citrin wurde von
5 beiden Plasmiden ausweislich der durch den spezifischen HRP-gekoppelten Sekundärantikörper gefärbten Banden erfolgreich exprimiert. Es ist jedoch bereits anhand der Bandenstärke deutlich erkennbar, dass das Plasmid mit der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:11 zu einer signifikant
10 gesteigerten Proteinausbeute im Vergleich zum Plasmid mit der herkömmlich optimierten Nukleotidsequenz SEQ ID NO:10 führte.

In dem in Fig. 6B gezeigten Säulendiagramm ist die relative Proteinausbeute an Citrin in Abhängigkeit von der jeweils
15 verwendeten Nukleotidsequenz dargestellt, wobei die quantitativ ermittelte Proteinmenge jeweils mithilfe der Ladekontrolle GAPDH als interner Standard hinsichtlich der Zellmenge normalisiert und auf die Proteinmenge aus dem Referenzversuch mit SEQ ID NO:10 normiert wurde. Die schraffierten Säulen zeigen das
20 Ergebnis der Quantifizierung anhand der Bandenintensität des Western Blots. Die Quantifizierung ergab, dass die erfindungsgemäße Optimierung der Citrin-Nukleotidsequenz mit $n = 3$ die Proteinausbeute gegenüber der herkömmlich optimierten Sequenz mehr als verdreifachte.

25
In dem in Fig. 6C gezeigten Säulendiagramm ist die relative Citrin-Fluoreszenz in Abhängigkeit von der jeweils verwendeten Nukleotidsequenz gezeigt. Die Citrin-Fluoreszenz ist hierbei ein Maß für den Anteil an gelöstem, funktionalem Protein. Die
30 Fluoreszenz wurde wie für Fig. 6B beschrieben gegen die Zellmenge normalisiert und auf die Fluoreszenz aus dem Referenzversuch mit SEQ ID NO:10 normiert. Die Quantifizierung der Fluoreszenz ergab, dass die erfindungsgemäße Optimierung der

Citrin-Nukleotidsequenz mit $n = 3$ die Ausbeute an gelöstem, funktionalem Protein etwa um den Faktor 4,7 steigerte. In Zusammenschau mit dem Ergebnis des Western Blots wird somit deutlich, dass die erfindungsgemäße Optimierung der

5 Nukleotidsequenzen des fluoreszierenden Proteins Citrin zu deutlich mehr (51%) löslichem Protein führte als die herkömmliche Optimierung.

Eine derart signifikante Verbesserung der heterologen

10 Expressionsrate eines löslichen Proteins in Säugerzellkultur mithilfe des erfindungsgemäßen Verfahrens zur Optimierung der proteincodierenden Nukleotidsequenz im Vergleich zum aktuellen Stand der Technik ist für den Fachmann nicht zu erwarten gewesen. Damit ist die Überlegenheit des erfindungsgemäßen

15 Optimierungsverfahrens gegenüber etablierten Konzepten für eukaryotische Expressionssysteme deutlich belegt.

Vergleichsbeispiel 4: Heterologe Expression von Citrin in *H. sapiens* (HEK293) Zellkultur

20 In einem weiteren Vergleichsversuch wurden die beiden in Vergleichsbeispiel 3 verwendeten optimierten Nukleotidsequenzen von Citrin in HEK293 Zellen exprimiert. Im Übrigen erfolgte die Versuchsdurchführung im Wesentlichen wie in Vergleichsbeispiel 3

25 beschrieben.

Die Ergebnisse sind in Fig. 7 gezeigt. Fig. 7A zeigt eine Abbildung des Western Blots, gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für das FLAG getaggte Citrin sowie gefärbt

30 mit dem HRP-gekoppeltem Sekundärantikörper für die Ladekontrolle GAPDH. Bahn 1 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:10 (Referenz), Bahn 2 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID

NO:11, die erfindungsgemäß optimiert wurde. Citrin wurde von beiden Plasmiden ausweislich der durch den spezifischen HRP-gekoppelten Sekundärantikörper gefärbten Banden auch in HEK293 Zellen erfolgreich exprimiert. Es ist bereits anhand der

5 Bandenstärke deutlich erkennbar, dass das Plasmid mit der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:11 zu einer signifikant gesteigerten Proteinausbeute im Vergleich zum Plasmid mit der herkömmlich optimierten Nukleotidsequenz SEQ ID NO:10 führte.

10

In dem in Fig. 7B gezeigten Säulendiagramm ist die relative Proteinausbeute an Citrin in Abhängigkeit von der jeweils verwendeten Nukleotidsequenz dargestellt, wobei die quantitativ ermittelte Proteinmenge jeweils mithilfe der Ladekontrolle GAPDH

15 als interner Standard hinsichtlich der Zellmenge normalisiert und auf die Proteinmenge aus dem Referenzversuch mit SEQ ID NO:10 normiert wurde. Die schraffierten Säulen zeigen das Ergebnis der Quantifizierung anhand der Bandenintensität des Western Blots. Die Quantifizierung ergab, dass die

20 erfindungsgemäße Optimierung der Citrin-Nukleotidsequenz mit $n = 3$ die Proteinausbeute um etwa 75% steigerte.

In dem in Fig. 7C gezeigten Säulendiagramm ist die relative Citrin-Fluoreszenz in Abhängigkeit von der jeweils verwendeten

25 Nukleotidsequenz gezeigt. Die Citrin-Fluoreszenz ist hierbei ein Maß für den Anteil an gelöstem, funktionalem Protein. Die Fluoreszenz wurde wie für Fig. 7B beschrieben gegen die Zellmenge normalisiert und auf die Fluoreszenz aus dem Referenzversuch mit SEQ ID NO:10 normiert. Die schraffierten

30 Säulen zeigen das Ergebnis der Quantifizierung anhand der Citrin-Fluoreszenz, welche funktionales und lösliches Protein wiedergibt. Die Quantifizierung ergab, dass die erfindungsgemäße Optimierung der Citrin-Nukleotidsequenz mit $n = 3$ die lösliche

Proteinausbeute etwa um den Faktor 2,3 steigerte. In
Zusammenschau mit dem Ergebnis des Western Blots wird somit
deutlich, dass die erfindungsgemäße Optimierung der
Nukleotidsequenzen des fluoreszierenden Proteins Citrin zu
5 deutlich mehr (32%) löslichem Protein führte als die
herkömmliche Optimierung.

Eine derart signifikante Verbesserung der heterologen
Expressionsrate in HEK293 Zellen mithilfe des erfindungsgemäßen
10 Verfahrens zur Optimierung der proteincodierenden
Nukleotidsequenz bestätigt das Ergebnis aus Vergleichsbeispiel 3
in einem weiteren Zielorganismus und impliziert eine universelle
Anwendbarkeit des erfindungsgemäßen Verfahrens in
Säugerzellkultur. Damit ist die Überlegenheit des
15 erfindungsgemäßen Optimierungsverfahrens gegenüber etablierten
Konzepten für eukaryotische Expressionssysteme evident.

Vergleichsbeispiel 5: Expression von *H. sapiens* STING1 ER exit
protein 1 (STEEP1) in HeLa Zellkultur

20 In einem weiteren Vergleichsversuch wurde die Nukleotidsequenz
des *H. sapiens* Proteins STEEP1 für die Expression in HeLa Zellen
nach dem erfindungsgemäßen Verfahren und in konventioneller
Weise gemäß Vergleichsbeispiel 3 als Referenz optimiert. STEEP1
25 ist ein humanes Protein, das in der Membran des
endoplasmatischen Retikulums gefunden wurde. Mutationen in
STEEP1 sind für mehrere Krankheitsbilder verantwortlich.
Proteine aus *H. sapiens* lassen sich im Allgemeinen schwer in
einem Wirtssystem exprimieren.

30 Für die Expression von STEEP1 in HeLa Zellen als Zielorganismus
wurden zwei pTwist CMV Expressionsplasmide von der Firma Twist
Bioscience bezogen, welche die Aminosäuresequenz von STEEP1 mit

einem FLAG-Tag gefolgt von einem doppelten Strep-Tag am N-Terminus unter einem konstitutiven Cytomegalovirus-Promotor codieren. N-terminal wurde die Aminosäuresequenz mit Met (Startcodon) und der Aminosäure Ala ergänzt. Vor dem Startcodon
5 wurde eine Kozak-Sequenz eingefügt. Eines der Plasmide enthielt die STEEP1-codierende Nukleotidsequenz mit FLAG- und doppeltem Strep-Tag nach konventioneller Optimierung durch den Hersteller (SEQ ID NO:13). Das andere Plasmid enthielt die STEEP1-codierende Nukleotidsequenz mit FLAG- und doppeltem Strep-Tag
10 nach Optimierung gemäß dem erfindungsgemäßen Verfahren mit $n = 3$ (SEQ ID NO:14), wobei für die erfindungsgemäße Bestimmung der relativen Codon-n-Tupel-Häufigkeit die für Proteine codierenden Teile des *H. sapiens* Genoms zugrunde gelegt wurden.

15 Im Übrigen erfolgte die Versuchsdurchführung für die erfindungsgemäße Sequenzoptimierung im Wesentlichen wie in Vergleichsbeispiel 3 beschrieben, ausgehend von der STEEP1 Wildtyp-Sequenz aus *H. sapiens* (SEQ ID NO:12) und unter Berücksichtigung der in Vergleichsbeispiel 3 genannten
20 unerwünschten Sequenzen, deren jeweilige Anzahl mit $r = 0,60$ gewichtet wurde. Die Anzahl L der in die Optimierung einbezogenen Codons in der Nukleotidsequenz ist in diesem Beispiel 762, die Kozak Sequenz am 5'-Ende wurde nicht optimiert.

25 Das Sequenzprotokoll zeigt erhebliche Unterschiede zwischen der Nukleotidsequenz vom Wildtyp (SEQ ID NO:12) und der Nukleotidsequenz, die gemäß dem erfindungsgemäßen Verfahren optimiert wurden (SEQ ID NO:14). So hat die Wildtypsequenz nur
30 81,6% identische Nukleotide mit der optimierten Sequenz aus dem erfindungsgemäßen Verfahren in der Ausführung mit $n = 3$.

Es ist weiterhin aus dem Sequenzprotoll unmittelbar zu entnehmen, dass sich die erfindungsgemäß optimierte Nukleotidsequenz (SEQ ID NO:14) von der gemäß dem Stand der Technik optimierten Sequenz (SEQ ID NO:13) auf Nukleotidebene
5 ebenfalls deutlich unterscheidet. So hat die nach dem Stand der Technik optimierte Sequenz (SEQ ID NO:13) nur 77,3% identische Nukleotide mit der gemäß dem erfindungsgemäßen Verfahren in der Ausführungsform mit $n = 3$ optimierten Sequenz (SEQ ID NO:14). Diese großen Unterschiede sind angesichts der vorgegebenen
10 Aminosäuresequenz aus fachmännischer Sicht überraschend und zeigen, dass das erfindungsgemäße Optimierungsverfahren auch für eine Optimierung in Säugerzellen, hier *Homo sapiens*, zu einem völlig anderen Ergebnis führt als das Verfahren gemäß dem aktuellen Stand der Technik.

15 Nach der Optimierung betrug der gewichtete Mittelwert in der erfindungsgemäßen Ausführungsform mit $n = 3$ (SEQ ID NO:14) $F_w = -9,8$. Eine analoge Vergleichsberechnung ergab einen gewichteten Mittelwert der relativen Codon-3-Tupel-Häufigkeiten
20 im Wildtyp (SEQ ID NO:12) von $F_w = -36,9$ und für die Referenzsequenz (SEQ ID NO:13) von $F_w = -87,1$.

Die Durchführung der Expressionsversuche erfolgte ebenfalls im Wesentlichen wie in Vergleichsbeispiel 3 beschrieben. Die
25 Ergebnisse sind in Fig. 8 gezeigt. Fig. 8A zeigt eine Abbildung des Western Blots, gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für das FLAG getaggte STEEP1 sowie gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für die Ladekontrolle GAPDH. Bahn 1 enthält das Expressionsprodukt des Plasmids mit
30 der eingefügten SEQ ID NO:13 (Referenz), Bahn 2 enthält das Expressionsprodukt des Plasmids mit der erfindungsgemäß optimierten SEQ ID NO:14. STEEP1 wurde von beiden Plasmiden ausweislich der durch den spezifischen HRP-gekoppelten

Sekundärantikörper gefärbten Banden erfolgreich exprimiert. Es ist jedoch bereits anhand der Bandenstärke deutlich erkennbar, dass das Plasmid mit der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:14 zu einer signifikant gesteigerten Proteinausbeute im Vergleich zum Plasmid mit der herkömmlich optimierten Nukleotidsequenz SEQ ID NO:13 führte.

In dem in Fig. 8B gezeigten Säulendiagramm ist die relative Proteinausbeute an STEEP1 in Abhängigkeit von der jeweils verwendeten Nukleotidsequenz dargestellt, wobei die quantitativ ermittelte Proteinmenge wie vorhergehend beschrieben mithilfe der Ladekontrolle auf die Zellmenge normalisiert und auf die Proteinmenge aus dem Referenzversuch mit SEQ ID NO:13 bezogen wurde. Die schraffierten Säulen zeigen das Ergebnis der Quantifizierung anhand der Bandenintensität des Western Blots. Die Quantifizierung ergab, dass die erfindungsgemäße Optimierung der STEEP1-Nukleotidsequenz mit $n = 3$ die Proteinausbeute um mehr als 75% gegenüber der herkömmlichen Optimierung steigerte.

Eine derart signifikante Verbesserung der Expressionsrate eines membranständigen Humanproteins mithilfe des erfindungsgemäßen Verfahrens zur Optimierung der proteincodierenden Nukleotidsequenz im Vergleich zum aktuellen Stand der Technik ist für den Fachmann nicht zu erwarten gewesen. Damit ist die Überlegenheit des erfindungsgemäßen Optimierungsverfahrens gegenüber etablierten Konzepten für eukaryotische Expressionssysteme deutlich belegt.

Vergleichsbeispiel 6: Expression von *H. sapiens* STING1 ER exit protein 1 (STEEP1) in HEK293 Zellkultur

In einem weiteren Vergleichsversuch wurden die beiden in Vergleichsbeispiel 5 verwendeten optimierten Nukleotidsequenzen

von STEEP1 in HEK293 Zellen exprimiert. Im Übrigen erfolgte die Versuchsdurchführung im Wesentlichen wie in Vergleichsbeispiel 5 beschrieben.

5 Die Ergebnisse sind in Fig. 9 gezeigt. Fig. 9A zeigt eine Abbildung des Western Blots, gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für das FLAG-getaggte STEEP1 sowie gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für die Ladekontrolle GAPDH. Bahn 1 enthält das Expressionsprodukt des Plasmids mit
10 der eingefügten SEQ ID NO:13 (Referenz), Bahn 2 enthält das Expressionsprodukt des Plasmids mit der eingefügten erfindungsgemäß optimierten SEQ ID NO:14. STEEP1 wurde von beiden Plasmiden ausweislich der durch den spezifischen HRP-gekoppelten Sekundärantikörper gefärbten Banden auch in HEK293
15 Zellen erfolgreich exprimiert. Es ist bereits anhand der Bandenstärke erkennbar, dass das Plasmid mit der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:14 zu einer signifikant gesteigerten Proteinausbeute im Vergleich zum Plasmid mit der herkömmlich optimierten Nukleotidsequenz SEQ ID NO:13 führte.

20

In dem in Fig. 9B gezeigten Säulendiagramm ist die relative Proteinausbeute an STEEP1 in Abhängigkeit von der jeweils verwendeten Nukleotidsequenz dargestellt, wobei die quantitativ ermittelte Proteinmenge wie vorhergehend beschrieben mithilfe
25 der Ladekontrolle auf die Zellmenge normalisiert und auf die Proteinmenge aus dem Referenzversuch mit SEQ ID NO:13 bezogen wurde. Die schraffierten Säulen zeigen das Ergebnis der Quantifizierung anhand der Bandenintensität des Western Blots. Die Quantifizierung ergab, dass die erfindungsgemäße Optimierung
30 der STEEP1-Nukleotidsequenz mit $n = 3$ die Proteinausbeute um 22% gegenüber der Expression der herkömmlich optimierten Sequenz steigerte.

Diese Ergebnisse bestätigen eine signifikante Verbesserung der Expressionsrate von STEEP1 in HEK293 Zellen mithilfe des erfindungsgemäßen Verfahrens zur Optimierung der proteincodierenden Nukleotidsequenz und untermauern die universelle Anwendbarkeit des erfindungsgemäßen Verfahrens für die Proteinexpression in Säugerzellen. Damit ist die Überlegenheit des erfindungsgemäßen Optimierungsverfahrens gegenüber etablierten Konzepten für eukaryotische Expressionssysteme erneut belegt.

Vergleichsbeispiel 7: Expression von *H. sapiens* Nitric oxide synthase-interacting protein (NOSIP) in HeLa Zellkultur

In einem weiteren Vergleichsversuch zu eukaryotischen Expressionssystemen wurde die Nukleotidsequenz des *H. sapiens* Proteins NOSIP für die Expression in HeLa nach dem erfindungsgemäßen Verfahren und nach dem Stand der Technik als Referenz optimiert. Die Referenzoptimierungen erfolgten in einer Variante gemäß Vergleichsbeispiel 3 und in einer zweiten Variante gemäß WO 2020/024917 A1. NOSIP moduliert die Aktivität und Lokalisation der Nitritoxid-Synthase und reguliert so die Nitritoxidproduktion, was entscheidend für die Entwicklung des menschlichen Hirns, des Auges und des Gesichts ist.

Für die Expression von NOSIP in HeLa Zellen als Zielorganismus wurden drei pTwist CMV Expressionsplasmide von der Firma Twist Bioscience bezogen, welche die Aminosäuresequenz von NOSIP mit einem FLAG-Tag gefolgt von einem doppelten Strep-Tag am N-Terminus unter einem konstitutiven Cytomegalovirus-Promotor codieren. N-terminal wurde die Aminosäuresequenz mit Met (Startcodon) und der Aminosäure Ala ergänzt. Vor dem Startcodon wurde eine Kozak-Sequenz eingefügt. Eines der Plasmide enthielt die NOSIP-codierende Nukleotidsequenz mit FLAG- und doppeltem

Strep-Tag nach Optimierung gemäß dem Stand der Technik durch den Optimierungsservice des Herstellers als Referenz (SEQ ID NO:16). Das zweite Plasmid enthielt die NOSIP-codierende Nukleotidsequenz mit FLAG- und doppeltem Strep-Tag nach
5 Optimierung gemäß WO 2020/024917 A1 als weitere Referenz (SEQ ID NO:17). Das dritte Plasmid enthielt die NOSIP-codierende Nukleotidsequenz mit FLAG- und doppeltem Strep-Tag nach Optimierung gemäß dem erfindungsgemäßen Verfahren mit $n = 3$ (SEQ ID NO:18), wobei für die erfindungsgemäße Bestimmung der
10 relativen Codon-n-Tupel-Häufigkeit die für Proteine codierenden Teile des *H. sapiens* Genoms zugrunde gelegt wurden.

Im Übrigen erfolgte die Versuchsdurchführung für die erfindungsgemäße Sequenzoptimierung im Wesentlichen wie in
15 Vergleichsbeispiel 3 beschrieben, ausgehend von der NOSIP Wildtyp-Sequenz aus *H. sapiens* (SEQ ID NO:15) unter Berücksichtigung der in Vergleichsbeispiel 3 genannten unerwünschten Sequenzen, deren jeweilige Anzahl mit $r = 0,64$ gewichtet wurde. Die Anzahl L der in die Optimierung
20 einbezogenen Codons in der Nukleotidsequenz ist in diesem Beispiel 1008, die Kozak Sequenz am 5'-Ende wurde nicht optimiert.

Das Sequenzprotokoll verdeutlicht auch hier, dass bereits auf
25 Nukleotidebene erhebliche Unterschiede zwischen der Nukleotidsequenz vom Wildtyp und der Nukleotidsequenz, die gemäß dem erfindungsgemäßen Verfahren optimiert wurde, bestehen. So weist die Wildtypsequenz (SEQ ID NO:15) nur 86,0% identische Nukleotide mit der optimierten Sequenz aus dem erfindungsgemäßen
30 Verfahren in der Ausführung mit $n = 3$ (SEQ ID NO:18) auf.

Es ist weiterhin aus dem Sequenzprotokoll zu erkennen, dass sich die erfindungsgemäß optimierte Nukleotidsequenz (SEQ ID NO:18)

auch von den gemäß dem Stand der Technik optimierten Sequenzen (SEQ ID NO:16, SEQ ID NO:17) auf Nukleotidebene deutlich unterscheidet. So hat die nach dem Stand der Technik optimierte Sequenz (SEQ ID NO:16) nur 76,3% identische Nukleotide mit der
5 gemäß dem erfindungsgemäßen Verfahren in der Ausführungsform mit $n = 3$ optimierten Sequenz (SEQ ID NO:18). Die nach dem Stand der Technik gemäß WO 2020/024917 A1 optimierte Sequenz (SEQ ID NO:17) weist nur 85,2% identische Nukleotide mit der gemäß dem erfindungsgemäßen Verfahren in der Ausführungsform mit $n = 3$
10 optimierten Sequenz (SEQ ID NO:18) auf. Diese großen Unterschiede zeigen erneut, dass das erfindungsgemäße Optimierungsverfahren auch für eine Optimierung in Säugerzellen, hier *Homo sapiens*, zu einem völlig anderen Ergebnis führt als das Verfahren gemäß dem aktuellen Stand der Technik.

15 Nach der Optimierung betrug der gewichtete Mittelwert in der erfindungsgemäßen Ausführungsform mit $n = 3$ (SEQ ID NO:18) $F_w = -13,1$. Eine analoge Vergleichsberechnung ergab einen gewichteten Mittelwert der relativen Codon-3-Tupel-Häufigkeiten
20 im Wildtyp (SEQ ID NO:15) $F_w = -39,1$, für die Referenzsequenz SEQ ID NO:16 $F_w = -148,0$ und für die Referenzsequenz SEQ ID NO:17 $F_w = -36,3$.

Im Übrigen erfolgte die Durchführung der Expressionsversuche im
25 Wesentlichen wie in Vergleichsbeispiel 3 beschrieben. Die Ergebnisse sind in Fig. 10 und Fig. 11 gezeigt.

Fig. 10 zeigt den Vergleich zwischen SEQ ID NO:16 und SEQ ID NO:18. Fig. 10A zeigt eine Abbildung des Western Blots, gefärbt
30 mit dem HRP-gekoppeltem Sekundärantikörper für das FLAG getaggte NOSIP sowie gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für die Ladekontrolle GAPDH. Bahn 1 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:16

(Referenz), Bahn 2 enthält das Expressionsprodukt des Plasmids mit der eingefügten SEQ ID NO:18, die erfindungsgemäß optimiert wurde. NOSIP wurde von beiden Plasmiden ausweislich der durch den spezifischen HRP-gekoppelten Sekundärantikörper gefärbten

5 Banden erfolgreich exprimiert, jedoch zeigt die Bandenstärke eine erheblich bessere Proteinausbeute durch die Expression der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:18 im Vergleich zum Plasmid mit der herkömmlich optimierten Nukleotidsequenz SEQ ID NO:16.

10

In dem in Fig. 10B gezeigten Säulendiagramm ist die relative Proteinausbeute an NOSIP in Abhängigkeit von der jeweils verwendeten Nukleotidsequenz dargestellt, wobei die quantitativ ermittelte Proteinmenge wie vorhergehend beschrieben mithilfe

15 der Ladekontrolle GAPDH zunächst auf die Zellmenge normalisiert und anschließend auf die Proteinmenge aus dem Referenzversuch mit SEQ ID NO:16 bezogen wurde. Die schraffierten Säulen zeigen das Ergebnis der Quantifizierung anhand der Bandenintensität des Western Blots. Die Quantifizierung ergab, dass durch die

20 erfindungsgemäße Optimierung der NOSIP-Nukleotidsequenz mit $n = 3$ die Proteinausbeute gegenüber der herkömmlichen Optimierung um den Faktor 31 erhöht werden konnte.

Eine derart signifikante Verbesserung der Expressionsrate eines

25 membranständigen Humanproteins mithilfe des erfindungsgemäßen Verfahrens zur Optimierung der proteincodierenden Nukleotidsequenz im Vergleich zum aktuellen Stand der Technik ist für den Fachmann bei Weitem nicht zu erwarten gewesen. Damit ist die Überlegenheit des erfindungsgemäßen

30 Optimierungsverfahrens gegenüber etablierten Konzepten für eukaryotische Expressionssysteme abermals deutlich belegt.

Fig. 11 zeigt den Vergleich zwischen SEQ ID NO:17 und SEQ ID NO:18. Fig. 11A zeigt eine Abbildung des Western Blots, gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für das FLAG getaggte NOSIP sowie gefärbt mit dem HRP-gekoppeltem Sekundärantikörper für die Ladekontrolle GAPDH. Bahn 1 enthält das Expressionsprodukt des Plasmids mit der gemäß WO 2020/024917 A1 optimierten Referenzsequenz SEQ ID NO:17, Bahn 2 enthält das Expressionsprodukt des Plasmids mit der erfindungsgemäß optimierten Sequenz SEQ ID NO:18. NOSIP wurde von beiden Plasmiden ausweislich der durch den spezifischen HRP-gekoppelten Sekundärantikörper gefärbten Banden erfolgreich exprimiert. Es ist bereits anhand der Bandenstärke erkennbar, dass das Plasmid mit der erfindungsgemäß optimierten Nukleotidsequenz SEQ ID NO:18 zu einer gesteigerten Proteinausbeute im Vergleich zum Plasmid mit der herkömmlich optimierten Nukleotidsequenz SEQ ID NO:17 führte.

In dem in Fig. 11B gezeigten Säulendiagramm ist die relative Proteinausbeute an NOSIP in Abhängigkeit von der jeweils verwendeten Nukleotidsequenz dargestellt, wobei die quantitativ ermittelte Proteinmenge wie vorhergehend beschrieben mithilfe der Ladekontrolle GAPDH zunächst auf die Zellmenge normalisiert und anschließend auf die Proteinmenge aus dem Referenzversuch mit SEQ ID NO:17 bezogen wurde. Die schraffierten Säulen zeigen das Ergebnis der Quantifizierung anhand der Bandenintensität des Western Blots. Die Quantifizierung ergab, dass die erfindungsgemäße Optimierung der NOSIP-Nukleotidsequenz mit n = 3 die Proteinausbeute um mehr als 60% steigerte.

Damit ist die Überlegenheit des erfindungsgemäßen Optimierungsverfahrens gegenüber der technischen Lehre des aktuellen Standes der Technik, hier WO 2020/024917 A1, auch für eukaryotische Expressionssysteme belegt.

Vergleichsbeispiel 8: Expression von NOSIP in HEK293 Zellkultur

In einem weiteren Vergleichsversuch wurden die beiden in
5 Vergleichsbeispiel 7 verwendeten optimierten Nukleotidsequenzen
von NOSIP SEQ ID NO:16 (Referenz) und SEQ ID NO:18 in HEK293
Zellen exprimiert. Im Übrigen erfolgte die Versuchsdurchführung
im Wesentlichen wie in Vergleichsbeispiel 7 beschrieben.

10 Die Ergebnisse sind in Fig. 12 gezeigt. Fig. 12A zeigt eine
Abbildung des Western Blots, gefärbt mit dem HRP-gekoppeltem
Sekundärantikörper für das FLAG getaggte NOSIP sowie gefärbt mit
dem HRP-gekoppeltem Sekundärantikörper für die Ladekontrolle
GAPDH. Bahn 1 enthält das Expressionsprodukt des Plasmids mit
15 der konventionell optimierten SEQ ID NO:16 (Referenz), Bahn 2
enthält das Expressionsprodukt des Plasmids mit der
erfindungsgemäß optimierten SEQ ID NO:18. NOSIP ließ sich mit
beiden Plasmiden ausweislich der durch den spezifischen HRP-
gekoppelten Sekundärantikörper gefärbten Banden auch in HEK293
20 Zellen erfolgreich exprimieren, wobei die Bandenstärke eine
deutlich bessere Expression der erfindungsgemäß optimierten
Nukleotidsequenz SEQ ID NO:18 zeigt.

In dem in Fig. 12B gezeigten Säulendiagramm ist die relative
25 Proteinausbeute an NOSIP in Abhängigkeit von der jeweils
verwendeten Nukleotidsequenz dargestellt, wobei die quantitativ
ermittelte Proteinmenge wie vorhergehend beschrieben mithilfe
der Ladekontrolle GAPDH zunächst auf die Zellmenge normalisiert
und anschließend auf die Proteinmenge aus dem Referenzversuch
30 mit SEQ ID NO:16 bezogen wurde. Die schraffierten Säulen zeigen
das Ergebnis der Quantifizierung anhand der Bandenintensität des
Western Blots. Die Quantifizierung ergab, dass die
erfindungsgemäße Optimierung der NOSIP-Nukleotidsequenz mit $n =$

3 die Proteinausbeute auch im HEK293 Expressionssystem gegenüber der herkömmlichen Sequenzoptimierung um etwa das Vierfache steigerte.

5 Auch hier ist wie im Falle von Vergleichsversuchen 4 und 6 eine signifikante Verbesserung der Expressionsrate in HEK293 Zellen mithilfe des erfindungsgemäßen Verfahrens zur Optimierung der proteincodierenden Nukleotidsequenz bestätigt. Eine universelle Anwendbarkeit des erfindungsgemäßen Verfahrens in
10 Säugerzellkultur ist dadurch gegeben. Damit ist die Überlegenheit des erfindungsgemäßen Optimierungsverfahrens gegenüber etablierten Optimierungskonzepten für eukaryotische Expressionssysteme erneut bestätigt.

15 Vergleichsbeispiel 9: Heterologe Expression von eqFP611 im Cyanobakterium *Synechococcus elongatus*

In einem weiteren Vergleichsversuch wurden die Nukleotidsequenzen des fluoreszierenden Proteins eqFP611 für die
20 heterologe Expression in *S. elongatus* nach dem erfindungsgemäßen Verfahren und nach dem Verfahren gemäß WO 2020/024917 A1 als Referenz optimiert. eqFP611 ist ein rot fluoreszierendes Protein (RFP) aus *Entacmaea quadricolor* und wird häufig für Reporter-Assays und in der Fluoreszenzmikroskopie verwendet.

25 Für die Expression von eqFP611 in *S. elongatus* als Zielorganismus wurden zwei pSyn-6 Expressionsplasmide (Thermo Fisher, Waltham, MA, USA) von der Firma Genscript bezogen, welche die Aminosäuresequenz von eqFP611 mit einem doppelten
30 Strep-Tag am N-Terminus unter einem konstitutivem psbA1 Promotor codiert. Eines der Plasmide enthielt die eqFP611-codierende Nukleotidsequenz mit doppeltem Strep-Tag nach Optimierung gemäß WO 2020/024917 A1 (SEQ ID NO:20). Das andere Plasmid enthielt

die eqFP611-codierende Nukleotidsequenz mit doppeltem Strep-Tag nach Optimierung gemäß dem erfindungsgemäßen Verfahren mit $n = 3$ (SEQ ID NO:21), wobei für die erfindungsgemäße Bestimmung der relativen Codon-n-Tupel-Häufigkeit die für Proteine codierenden Teile des *S. elongatus* Genoms zugrunde gelegt wurden.

Das Optimierungsverfahren erfolgte wie in Vergleichsbeispiel 1 beschrieben. Analog zur Optimierung für *E. coli* als Wirt wurden für *S. elongatus* als optionaler Schritt in Feld 108 die unerwünschten Sequenzmotive wie z. B. die ribosomale Bindestelle AGGAGG, von denen dem Fachmann bekannt ist, dass sie die Expression in *S. elongatus* beeinträchtigen können, eingegeben. Weitere unerwünschte Sequenzmotive waren GGGGGG, CGCGCG, ATTTA, CATATG, GGTACC und GTCGAC. Die Anzahl der unerwünschten Sequenzmotive wurde mit $r = 0,62$ gewichtet. Die Anzahl L der in die Optimierung einbezogenen Codons in der Nukleotidsequenz ist in diesem Beispiel 768, die Terminatorsequenz und die terminalen Restriktionsschnittstellen wurden nicht optimiert.

Das Sequenzprotokoll zeigt, dass sich die für die heterologe Expression von eqFP611 in *S. elongatus* erfindungsgemäß optimierte Nukleotidsequenz (SEQ ID NO:21) von der gemäß dem Stand der Technik optimierten Sequenz (SEQ ID NO:20) auf Nukleotidebene deutlich unterscheidet. So hat die nach dem Stand der Technik optimierte Sequenz (SEQ ID NO:20) nur 89,8% identische Nukleotide mit der gemäß dem erfindungsgemäßen Verfahren in der Ausführungsform mit $n = 3$ optimierten Sequenz (SEQ ID NO:21). Diese großen Unterschiede sind angesichts der vorgegebenen Aminosäuresequenz aus fachmännischer Sicht überraschend und zeigen, dass das erfindungsgemäße Optimierungsverfahren auch für eine Optimierung in Cyanobakterien, hier *S. elongatus*, zu einem völlig anderen

Ergebnis führt als das Optimierungsverfahren gemäß dem aktuellen Stand der Technik.

Nach der Optimierung betrug der gewichtete Mittelwert in der
5 erfindungsgemäßen Ausführungsform mit $n = 3$ (SEQ ID NO:21)
 $F_w = -8,7$. Eine analoge Vergleichsberechnung ergab einen
gewichteten Mittelwert der relativen Codon-3-Tupel-Häufigkeiten
für die Referenzsequenz gemäß WO 2020/024917 A1 (SEQ ID NO:20)
von $F_w = -496,2$.

10

Für die Expression von eqFP611 in *S. elongatus* wurde das
Herstellerprotokoll des „GeneArt algal protein expression
system“ (Thermo Fisher) befolgt. Die Weiterbehandlung der
cyanobakteriellen Biomasse erfolgt wie in Vergleichsbeispiel 1
15 für *E. coli* beschrieben. Das exprimierte eqFP611 Protein wird
über Streptactin gereinigt und durch SDS-PAGE quantifiziert.
Zusätzlich werden Fraktionen der eqFP611-Zellysate für zwei
Minuten bei 13.000 g zentrifugiert und die eqFP611-Fluoreszenz
im Überstand in einem Tecan Spark Plate Reader bei einer
20 Anregungswellenlänge von 559 nm und einer Emissionswellenlänge
von 611 nm in Triplikaten gemessen.

Als Ergebnis dieses Versuches ist auch für Cyanobakterien eine
ähnlich hohe Steigerung der löslichen Proteinausbeute durch das
25 erfindungsgemäße Verfahren zu erwarten, wie bereits in den
Vergleichsbeispielen 1-8 für bakterielle und eukaryotische
Expressionssysteme beschrieben.

Die Erfindung ist nicht durch die Beschreibung anhand der
30 Ausführungsbeispiele auf diese beschränkt. Vielmehr umfasst die
Erfindung jedes neue Merkmal sowie jede Kombination von
Merkmalen, was insbesondere jede Kombination von Merkmalen in
den Patentansprüchen und der Beschreibung beinhaltet, auch wenn

dieses Merkmal oder diese Kombination von Merkmalen selbst nicht explizit in den Patentansprüchen, der Beschreibung oder den Ausführungsbeispielen angegeben ist.

Patentansprüche

1. Ein Verfahren zur Optimierung einer Nukleotidsequenz für die Expression einer vorgegebenen Aminosäuresequenz in mindestens
5 einem Zielorganismus, wobei die Nukleotidsequenz eine Vielzahl von Basentripletts umfasst und an mindestens einer Änderungsposition der Nukleotidsequenz ein Basentriplett, das eine Aminosäure der vorgegebenen Aminosäuresequenz codiert, durch ein synonymes Basentriplett, das dieselbe Aminosäure
10 der vorgegebenen Aminosäuresequenz codiert, ersetzt wird, um die Nukleotidsequenz für die Expression in dem mindestens einen Zielorganismus zu optimieren,
- dadurch gekennzeichnet, dass** die mindestens eine Änderungsposition eine direkte Aufeinanderfolge von n
15 Basentripletts umfasst, die ein erstes Codon-n-Tupel bildet und einen Sequenzabschnitt von n Aminosäuren der vorgegebenen Aminosäuresequenz codiert, der ein Aminosäure-n-Tupel bildet, das mit einer vorbestimmten Menge von Aminosäure-n-Tupel-Ereignissen im Genom oder eines Teils davon des mindestens
20 einen Zielorganismus und/oder in Genomen oder Teilen davon von zur Infektion des mindestens einen Zielorganismus befähigten Viren codiert wird, und
- mindestens eines der n Basentripletts der direkten Aufeinanderfolge durch ein synonymes Basentriplett ersetzt
25 wird, das so gewählt ist, dass ein zweites Codon-n-Tupel resultiert, das bezogen auf die Menge von Aminosäure-n-Tupel-Ereignissen eine höhere relative Codon-n-Tupel-Häufigkeit in dem Genom oder des Teils davon des mindestens einen Zielorganismus und/oder in den Genomen oder der Teile davon
30 der zur Infektion des mindestens einen Zielorganismus befähigten Viren aufweist als das erste Codon-n-Tupel,

- 78 -

wobei n eine natürliche Zahl größer gleich zwei und kleiner gleich einer Gesamtzahl N der Aminosäuren der vorgegebenen Aminosäuresequenz ist.

- 5 2. Das Verfahren nach Anspruch 1, **dadurch gekennzeichnet, dass** n kleiner gleich 50, kleiner gleich 40, kleiner gleich 30, kleiner gleich 20 oder kleiner gleich 10 ist.
3. Das Verfahren nach einem der Ansprüche 1 oder 2, dadurch
10 gekennzeichnet, dass n ausgewählt ist aus der Gruppe bestehend aus $n = 2$, $n = 3$, $n = 4$, $n = 5$ und beliebigen Kombinationen davon.
4. Das Verfahren nach einem der Ansprüche 1 bis 3, **dadurch**
15 **gekennzeichnet, dass** die Menge von Aminosäure- n -Tupel-Ereignissen anhand von mehreren proteincodierenden Genen und/oder Proteinen des mindestens einen Zielorganismus und/oder der zur Infektion des mindestens einen Zielorganismus befähigten Viren bestimmt wird und/oder sich
20 daraus ergibt.
5. Das Verfahren nach einem der Ansprüche 1 bis 4, **dadurch**
gekennzeichnet, dass die relative Codon- n -Tupel-Häufigkeit anhand einer Menge von Ereignissen des jeweils ersten
25 und/oder zweiten Codon- n -Tupels in mehreren proteincodierenden Genen des mindestens einen Zielorganismus und/oder der zur Infektion des mindestens einen Zielorganismus befähigten Viren, der die Menge von Aminosäure- n -Tupel-Ereignissen zugrunde gelegt wird, bestimmt
30 wird und/oder sich daraus ergibt.
6. Das Verfahren nach einem der Ansprüche 1 bis 5, **dadurch gekennzeichnet, dass** die folgenden Schritte umfasst sind:

- a) Bestimmen der mindestens einen Änderungsposition;
 - b) Ersetzen des mindestens einen Basentriplets der mindestens einen Änderungsposition durch das synonyme Basentriplett;
 - 5 c) Ermitteln der relativen Codon-n-Tupel-Häufigkeit des resultierenden zweiten Codon-n-Tupels.
7. Das Verfahren nach einem der Ansprüche 1 bis 6, **dadurch gekennzeichnet, dass** die mindestens eine Änderungsposition
- 10 eine Vielzahl von Änderungspositionen ist, wobei in jeder der Änderungspositionen mindestens eines der n Basentriplets der direkten Aufeinanderfolge durch ein synonymes Basentriplett ersetzt wird, wobei die synonymen Basentriplets so gewählt sind, dass zumindest ein Teil der resultierenden zweiten
- 15 Codon-n-Tupel eine höhere relative Codon-n-Tupel-Häufigkeit aufweist als die jeweiligen ersten Codon-n-Tupel.
8. Das Verfahren nach Anspruch 7, **dadurch gekennzeichnet, dass** in der Änderungsposition mit dem ersten Codon-n-Tupel, das
- 20 die geringste relative Codon-n-Tupel-Häufigkeit aufweist, mindestens eines der n Basentriplets der direkten Aufeinanderfolge durch ein synonymes Basentriplett ersetzt wird, das so gewählt ist, dass das resultierende zweite Codon-n-Tupel eine höhere relative Codon-n-Tupel-Häufigkeit
- 25 als das erste Codon-n-Tupel aufweist.
9. Das Verfahren nach einem der Ansprüche 7 oder 8, **dadurch gekennzeichnet, dass** sich die direkten Aufeinanderfolgen der n Basentriplets von mindestens zwei Änderungspositionen
- 30 überschneiden, wobei das Basentriplett, das durch das synonyme Basentriplett ersetzt wird, von beiden Änderungspositionen gleichzeitig umfasst ist.

10. Das Verfahren nach Anspruch 9, **dadurch gekennzeichnet, dass**
das synonyme Basentriplett so gewählt ist, dass in einer der
beiden Änderungspositionen das resultierende zweite Codon-n-
Tupel eine geringere relative Codon-n-Tupel-Häufigkeit und in
5 der anderen der beiden Änderungspositionen das resultierende
zweite Codon-n-Tupel eine höhere relative Codon-n-Tupel-
Häufigkeit als das jeweilige erste Codon-n-Tupel aufweist.
11. Das Verfahren nach einem der Ansprüche 7 bis 10, **dadurch**
10 **gekennzeichnet, dass** die relative Codon-n-Tupel-Häufigkeit
der ersten Codon-n-Tupel und die relative Codon-n-Tupel-
Häufigkeit der zweiten Codon-n-Tupel in den
Änderungspositionen jeweils ein globales Minimum aufweist,
wobei das globale Minimum der zweiten Codon-n-Tupel größer
15 ist als das globale Minimum der ersten Codon-n-Tupel.
12. Das Verfahren nach einem der Ansprüche 7 bis 11, **dadurch**
gekennzeichnet, dass die synonymen Basentriplets so gewählt
werden, dass die relative Codon-n-Tupel-Häufigkeit der
20 zweiten Codon-n-Tupel einen größtmöglichen Mindestwert
erreicht oder nicht mehr als 50% unterhalb des größtmöglichen
Mindestwertes liegt.
13. Das Verfahren nach einem der Ansprüche 7 bis 12, **dadurch**
25 **gekennzeichnet, dass** die synonymen Basentriplets so gewählt
werden, dass ein Mittelwert aus den relativen Codon-n-Tupel-
Häufigkeiten der zweiten Codon-n-Tupel einen Höchstwert
erreicht oder nicht mehr als 50% unterhalb des erreichbaren
Höchstwertes liegt.
30
14. Das Verfahren nach Anspruch 13, **dadurch gekennzeichnet, dass**
der Mittelwert eine degressive Wichtung der relativen Codon-
n-Tupel-Häufigkeiten enthält, die so konfiguriert ist, dass

eine hohe relative Codon-n-Tupel-Häufigkeit im Vergleich zu einer niedrigeren relativen Codon-n-Tupel-Häufigkeit einen betragsmäßig unterproportionalen Einfluss auf den Mittelwert hat.

5

15. Das Verfahren nach einem der Ansprüche 7 bis 14, **dadurch gekennzeichnet, dass** sich mindestens zwei der Änderungspositionen in der Zahl n unterscheiden oder n für mindestens zwei der Änderungspositionen unterschiedlich gewählt wird.
- 10
16. Das Verfahren nach einem der Ansprüche 7 bis 15, **dadurch gekennzeichnet, dass** das Ersetzen der Basentriplets durch die synonymen Basentriplets in mehreren Iterationsschritten durch ein computergestütztes Optimierungsverfahren durchgeführt wird.
- 15
17. Das Verfahren nach Anspruch 16, **dadurch gekennzeichnet, dass** das computergestützte Optimierungsverfahren ein Approximationsverfahren und/oder simuliertes Abkühlungsverfahren (Simulated Annealing) umfasst.
- 20
18. Das Verfahren nach einem der Ansprüche 7 bis 17, **dadurch gekennzeichnet, dass** die Änderungspositionen zusammen mindestens 50% der Basentriplets der Nukleotidsequenz umfassen, die eine Aminosäure der vorgegebenen Aminosäuresequenz codieren.
- 25
19. Das Verfahren nach einem der Ansprüche 1 bis 18, **dadurch gekennzeichnet, dass** der mindestens eine Zielorganismus eine Mehrzahl von verschiedenen Zielorganismen ist, wobei die relative Codon-n-Tupel-Häufigkeit eine genomabhängige Wichtung enthält, die so konfiguriert ist, dass ein
- 30

Größenunterschied in den Genomen oder den Teilen davon der verschiedenen Zielorganismen zumindest teilweise kompensiert wird.

- 5 20. Das Verfahren nach einem der Ansprüche 1 bis 19, **dadurch gekennzeichnet, dass** nach der Optimierung der Nukleotidsequenz die in dem mindestens einen Zielorganismus exprimierte Aminosäuresequenz eine größere Löslichkeit aufweist und/oder zu einem größeren Anteil in gelöster Form
10 vorliegt als vor der Optimierung.
21. Ein Verfahren zur Optimierung einer Nukleotidsequenz für die Expression einer vorgegebenen Aminosäuresequenz in mindestens einem Zielorganismus, wobei die Nukleotidsequenz eine
15 Vielzahl von Basentripletts umfasst und an mindestens einer Änderungsposition der Nukleotidsequenz ein Basentriplett, das eine Aminosäure der vorgegebenen Aminosäuresequenz codiert, durch ein synonymes Basentriplett, das dieselbe Aminosäure der vorgegebenen Aminosäuresequenz codiert, ersetzt wird, um
20 die Nukleotidsequenz für die Expression in dem mindestens einen Zielorganismus zu optimieren,
dadurch gekennzeichnet, dass die mindestens eine Änderungsposition eine direkte Aufeinanderfolge von n Basentripletts umfasst, die ein erstes Codon-n-Tupel bildet
25 und einen Sequenzabschnitt von n Aminosäuren der vorgegebenen Aminosäuresequenz codiert, der ein Aminosäure-n-Tupel bildet, und
mindestens eines der n Basentripletts der direkten Aufeinanderfolge durch ein synonymes Basentriplett ersetzt
30 wird, welches anhand einer Schätzfunktion so ausgewählt wird, dass ein zweites Codon-n-Tupel resultiert, welches das Aminosäure-n-Tupel mit einer größeren Wahrscheinlichkeit im Genom oder eines Teils davon des mindestens einen

- Zielorganismus und/oder in Genomen oder Teilen davon von zur Infektion des mindestens einen Zielorganismus befähigten Viren codiert als das erste Codon-n-Tupel,
wobei n eine natürliche Zahl größer gleich zwei und
5 kleiner gleich einer Gesamtzahl N der Aminosäuren der vorgegebenen Aminosäuresequenz ist.
22. Eine Verwendung einer gemäß dem Verfahren nach einem der Ansprüche 1 bis 21 optimierten Nukleotidsequenz zur
10 Herstellung einer synthetischen DNA und/oder zur Proteinexpression in mindestens einem Zielorganismus.
23. Ein Computerprogramm mit Programmcodemitteln, **dadurch gekennzeichnet, dass** die Programmcodemittel zur Durchführung
15 eines Verfahrens nach einem der Ansprüche 1 bis 21 eingerichtet sind, wenn das Computerprogramm auf einem Rechner ausgeführt wird.
24. Ein computerlesbares Speichermedium mit einem in
20 computerlesbarer Form gespeicherten Computerprogramm nach Anspruch 23.
25. Eine Vorrichtung zur Optimierung und/oder Herstellung einer Nukleotidsequenz für die Expression einer vorgegebenen
25 Aminosäuresequenz in mindestens einem Zielorganismus, mit einer Recheneinrichtung, die zur Durchführung eines Verfahrens nach einem der Ansprüche 1 bis 21 eingerichtet ist.
- 30 26. Ein Verfahren zur Herstellung eines Proteins, umfassend Bereitstellen einer Nukleotidsequenz, die gemäß einem Verfahren nach einem der Ansprüche 1 bis 21 für die

Expression des Proteins in mindestens einem Zielorganismus optimiert wurde,

 Exprimieren des Proteins in dem mindestens einen Zielorganismus.

5

27. Ein Nukleinsäuremolekül, das eine Nukleotidsequenz umfasst, die durch ein Verfahren nach einem der Ansprüche 1 bis 21 erhalten wurde.

10 28. Ein Vektor, der ein Nukleinsäuremolekül nach Anspruch 27 umfasst.

29. Eine rekombinante Wirtszelle, die ein Nukleinsäuremolekül nach Anspruch 27 oder einen Vektor nach Anspruch 28 enthält.

15