KI-optimierte Codonnutzung

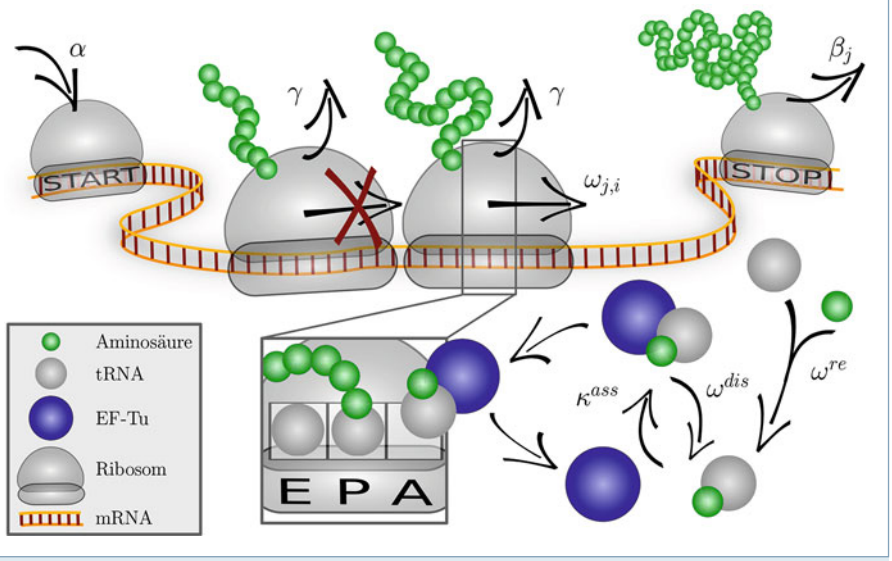
Modellentwicklung und maschinelles Lernen erhöhen die Proteinausbeute

JAN-HENDRIK TRÖSEMEIER[1], SOPHIA RUDORF[2], HOLGER LÖßNER[1], BENJAMIN HOFNER[1], CHRISTEL KAMP[1]
[1] PAUL-EHRLICH-INSTITUT, LANGEN
[2] MAX-PLANCK-INSTITUT FÜR KOLLOID- UND GRENZFLÄCHENFORSCHUNG, POTSDAM

Heterologous expression of genes requires their adaptation to the host organism to achieve adequate protein synthesis rates. Typically codons are adjusted to resemble those seen in highly expressed genes of the host organism which lacks a deeper understanding of codon optimality. The codon-specific elongation model (COSEM) identifies optimal codon choices by simulating ribosome dynamics during mRNA translation. COSEM is used in combination with machine learning techniques to predict protein abundance and to optimize codon usage.

DOI: 10.1007/s12268-020-1369-3

Proteine für biotechnologisch hergestellte Arzneimittel werden in Zellkulturen produziert. Hierzu werden Gene mit den Informationen über die Aminosäuresequenz der gewünschten Proteine in Bakterien- oder Säugerzellen integriert. Diese Zellen dienen als Synthesefabriken für die fremden Proteine. Für eine ausreichende Proteinproduktion ist häufig eine Anpassung der eingebrachten Gene an die Wirtszelle erforderlich: Die genetische Codierung der Aminosäuren ist nämlich nicht eindeutig. Jedes Triplett der Nukleobasen A, C, G und T der Boten-RNA (mRNA) codiert eine Aminosäure – den 64 möglichen Kombinationen aus drei Nukleotiden stehen 21 Aminosäuren gegenüber. Je nach Wirtsorganismus erweisen sich andere Codons als optimal, die besonders schnell oder genau ausgelesen werden (Translation). Dies ist eine Konsequenz der unterschiedlichen Ausstattung der Wirtszellen und erfordert eine optimierte Codonwahl [1].

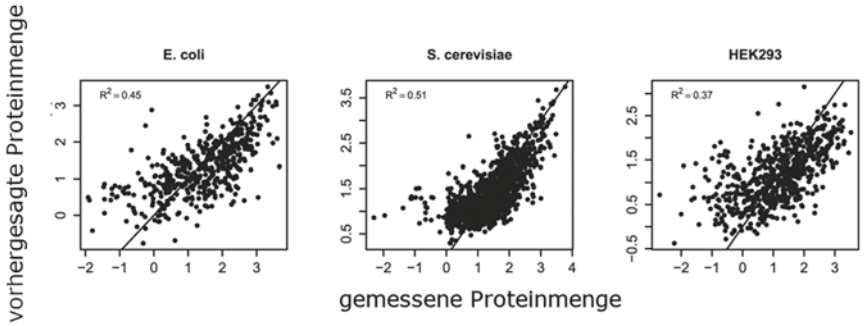


▲ **Abb. 1:** Codonspezifisches Elongationsmodell (COSEM) zur Bestimmung der Proteinsyntheserate. Nach dem Beginn der Translation (mit Rate α) erfolgt die Synthese mit codonspezifischen Raten $\omega_{j,i}$, die sich aus der Interaktion von Ribosomen, Transfer-RNAs (tRNAs) und Elongationsfaktoren (EF-Tu) ergeben (mit Raten κ^{ass}, ω^{dis} und ω^{re} siehe [4]). Die Synthese kann vollständig ausgeführt (mit Abschlussrate β_j) oder vorzeitig beendet (mit Rate γ) werden (adaptiert aus [2]).

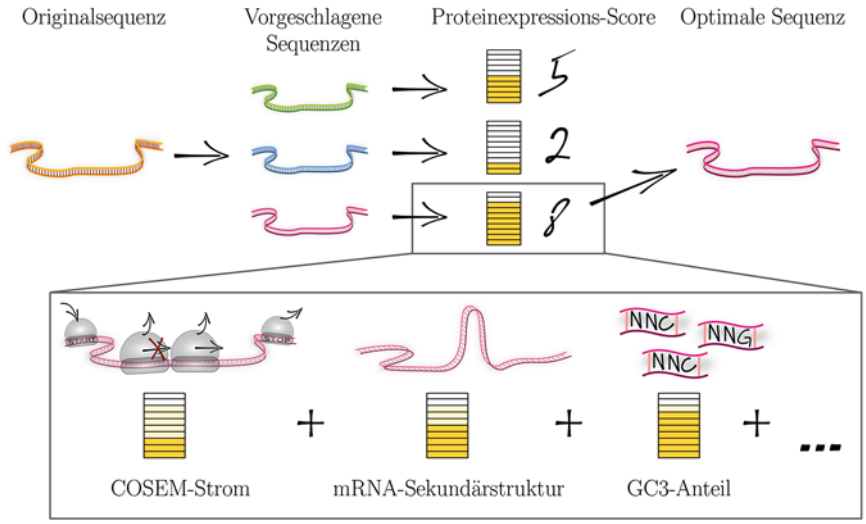
Gängige Optimierungsstrategien streben eine Nachahmung der Codonhäufigkeiten an, wie sie bei besonders oft abgelesenen, natürlichen Genen des Zielorganismus vorkommen. Diese heuristischen Verfahren liefern im Allgemeinen gute Ergebnisse, sind jedoch ohne direkten Bezug zur Funktionsweise der Translation und somit ohne Alternativstrategien für Fälle mit unerwartet schlechter Proteinausbeute. Im Gegensatz dazu bietet ein mechanistisches Modell der Proteinbiosynthese wie das codonspezifische Elongationsmodell (COSEM; **Abb. 1**) neuartige Einblicke in die zugrunde liegenden Funktionsweisen und entsprechend neue Ansätze zur Optimierung. Komplementär zu diesen mechanistischen Modellen können mit Methoden des maschinellen Lernens die wesentlichen Einflussfaktoren der Proteinproduktion ermittelt werden (**Abb. 3**). In der Kombination erhalten wir so ein qualitativ verbessertes Werkzeug zur Codonoptimierung [2, 3].

Wie viel Protein können wir erwarten?

Wir verwenden COSEM (**Abb. 1,** [2]) zur Simulation der Dynamik der Proteinsynthese und zur Vorhersage von Proteinsyntheseraten: Ribosomen binden mit einer festen Rate α an die mRNA und bewegen sich anschließend mit einer codonabhängigen Rate ($\omega_{j,i}$) [4] von Codon zu Codon, es sein denn, sie werden durch vorhergehende Ribosomen blockiert. Gelegentlich (mit Rate γ) kommt es zu einem vorzeitigen Abbruch der Proteinsynthese. Aus der kollektiven Bewegung der Ribosomen lässt sich in der Simu-



▲ **Abb. 2:** Vergleich von gemessener mit vorhergesagter Proteinausbeute in den Modellorganismen *Escherichia coli* und *Saccharomyces cerevisiae* und in HEK-293-Zellen. Das Bestimmtheitsmaß R^2 gibt dabei den Anteil der Varianz an, den das Modell beschreibt (adaptiert aus [2]).



▲ **Abb. 3:** Codonoptimierung. Aus verschiedenen vorgeschlagenen Sequenzen wird basierend auf nutzerdefinierten Sequenzeigenschaften (Box) der Proteinexpressions-Score bestimmt und mit diesem eine optimale Sequenz ausgewählt. Aufgrund der Modularität des Modells kann eine Sequenz hinsichtlich sehr unterschiedlicher, individueller Zielfunktionen flexibel optimiert werden. Betrachtete Sequenzeigenschaften sind unter anderem der COSEM-Strom (Proteinsyntheserate aus dem codonspezifischen Elongationsmodell, COSEM), die mRNA-Sekundärstruktur (Faltungsenergie der mRNA) und der GC3-Anteil (Anteil von Guanin und Cytosin in der mRNA) (adaptiert aus [2]).

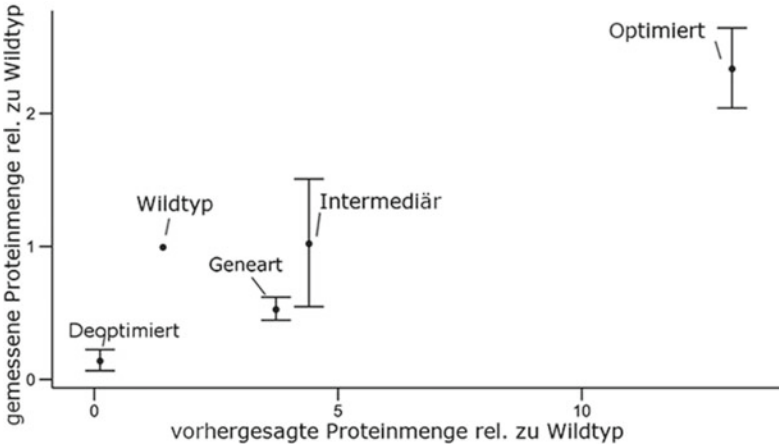
lation eine Proteinsyntheserate ableiten, die spezifisch ist für die Codonzusammensetzung der betrachteten mRNA.

Zur Vorhersage der Proteinmenge in der Wirtszelle setzen wir den Beitrag dieser simulierten Proteinsyntheserate (COSEM-Strom) in einen funktionalen Zusammenhang mit weiteren relevanten Vorhersagevariablen, wie z. B. der Sekundärstruktur der mRNA. Wir trainieren ein additives statistisches Modell [5] mit publizierten Daten zu Proteinmengen in *Escherichia coli*, *Saccharomyces cerevisiae* und menschlichen Zellen (HEK-293) und erhalten so einen Proteinexpressions-Score. Zusammenfassend erlauben uns Methoden des maschinellen Lernens, wesentliche Einflussfaktoren der Proteinproduktion und deren funktionalen Beitrag zur Proteinausbeute in Wirtsorganismen abzuschätzen. Mit dem resultierenden Modell können wir die Proteinausbeute auf Basis der Gensequenz vorhersagen (**Abb. 2**). In analoger Weise kann das Modell auch auf weitere Wirtsorganismen trainiert werden.

Welche Gensequenz optimiert die Proteinausbeute?

Der so etablierte Zusammenhang zwischen Proteinausbeute und Codonnutzung kann umgekehrt zur Optimierung der Proteinausbeute im Wirtsorganismus genutzt werden. So können wir gezielt nach Gensequenzen suchen, die die Proteinausbeute maximieren (gemessen über den Proteinexpressions-Score). Wir lösen dieses Optimierungsproblem über einen stochastischen Algorithmus: Dieser schlägt verschiedene mögliche Sequenzen vor und wählt schließlich die optimale Sequenz aus (**Abb. 3**). Dabei berücksichtigt der Algorithmus auch weitere Sequenzanforderungen, wie z. B. eine hohe Translationsgenauigkeit oder die Vermeidung bestimmter Sequenzmotive. Der modulare Ansatz unseres Modells erlaubt neben der Maximierung der Proteinausbeute auch die Festlegung von individuellen Optimierungszielen. So kann der Optimierungsmechanismus auch genutzt werden, um die Expression von Proteinen zu senken. Eine solche Deoptimierung der Codonnutzung kann beispielsweise eine genetische Abschwächung von Pathogenen bewirken, sodass sie als Impfstoff eingesetzt werden können.

Die beschriebene Methode der Codonoptimierung wendeten wir exemplarisch auf die Synthese von Ovalbumin in Salmonellen an. Ovalbumin ist ein relevantes Allergen aus Hühnereiweiß, dessen Produktion in verschiedenen Anwendungsfeldern von Bedeutung ist: Ovalbumin-spezifische Impfstrategien werden untersucht, um Eiweißallergien im Lebensmittelbereich zu mildern. Weiterhin ist Ovalbumin ein etabliertes Modellantigen für die Evaluierung von B- und T-Zell-vermittelten Immunantworten in Mäusen. Gleichzeitig führte eine Codonoptimierung mit einem Standardverfahren (GeneOptimizer der Firma Geneart/Life Technologies) zu keiner Steigerung der Proteinproduktion im Wirtsorganismus. Wir entwickelten und testeten verschiedene Ovalbumin-codierende mRNA-Sequenzen, für die unsere Methode eine geringe, eine intermediäre und eine hohe Proteinausbeute in Salmonellen vorhersagt. **Abbildung 4** zeigt, dass unsere Methode zu einer deutlichen Steigerung der Proteinausbeute gegenüber der etablierten Methode führt. Die Weiterentwicklung des mechanistischen Modells mithilfe des maschinellen Lernens ermöglicht darüber hinaus quantitative Aussagen über die zu erwartende Proteinausbeute, die mit bisherigen Codonoptimierungsverfahren nicht möglich sind. Unser Ansatz stellt somit einen echten Fortschritt in der *in silico*-Codonoptimierung mit weitreichenden Anwendungen in der synthetischen Biologie dar. Der Optimierungsalgorithmus ist in der Software



◀ **Abb. 4:** Vorhergesagte und gemessene Proteinmengen von in Salmonellen synthetisiertem Hühnereiweiß (Ovalbumin) für verschiedene Genvarianten (Mittelwert und Standardabweichung). Unsere Methode liefert eine höhere Proteinausbeute als das Standardverfahren (Geneart: optimiert mit GenOptimizer, Geneart/Life Technologies) und darüber hinaus quantitative Aussagen über die zu erwartende Proteinausbeute (je normiert auf den Wildtyp; optimiert auf niedrige [deoptimiert], intermediäre und hohe Proteinausbeute [optimiert]) (adaptiert aus [2]).

OCTOPOS (*optimized codon translation for protein synthesis*) umgesetzt und macht die oben beschriebene Optimierung von Genen für einen weiten Anwenderkreis zugänglich.

Danksagung

Wir danken unseren Kooperationspartnern und den Mitarbeitern des Paul-Ehrlich-Instituts sowie des Max-Planck-Instituts für Kolloid- und Grenzflächenforschung für die fruchtbare Zusammenarbeit, weiterhin der Adolf-Messer-Stiftung für die finanzielle Unterstützung des Projektes. ■

Literatur

[1] Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12:32–42
[2] Trösemeier J, Rudorf S, Loessner H et al. (2019) Optimizing the dynamics of protein expression. Sci Rep 9:7511
[3] International Patent Application No. PCT/EP2017/081685, Codonoptimierung, https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2018104385
[4] Rudorf S, Lipowsky R (2015) Protein synthesis in *E. coli*: dependence of codon-specific elongation on tRNA concentration and codon usage. PLoS One 10:1–22
[5] Hofner B, Hothorn T, Kneib T et al. (2011) A framework for unbiased model selection based on boosting. J Comput Graph Stat 20:956–971

Funding: Open Access funding provided by Projekt DEAL.

Korrespondenzadresse:
Dr. Christel Kamp
Paul-Ehrlich-Institut
Bundesinstitut für Impfstoffe und biomedizinische Arzneimittel
Paul-Ehrlich-Straße 51–59
D-63225 Langen
christel.kamp@pei.de

AUTOREN

Jan-Hendrik Trösemeier
2007–2012 Physikstudium an der Universität Göttingen und am Max-Planck-Institut für Dynamik und Selbstorganisation, Göttingen. Bis **2016** Promotion an der Universität Frankfurt a. M. und am Paul-Ehrlich-Institut, Langen, bei Prof. Dr. I. Koch. Seit **2016** Entwickler in der Intel Deutschland GmbH.

Sophia Rudorf
2004–2009 Physikstudium an der Universität Potsdam, der University of California, USA, und der LMU München. **2015** Promotion am Max-Planck-Institut für Kolloid- und Grenzflächenforschung, Potsdam, bei Prof. Dr. R. Lipowsky. Seither dort Gruppenleiterin in der Abteilung für Theorie und Bio-Systeme.

Holger Lößner
1990–1996 Biochemiestudium an der HU Berlin. **2003** Promotion am Max-Planck-Institut für Infektionsbiologie, Berlin. **2002–2008** Postdoc am Helmholtz-Zentrum für Infektionsforschung, Braunschweig. Seit **2008** Wissenschaftler und Assessor am Paul-Ehrlich-Institut, Langen.

Benjamin Hofner
2002–2008 Statistikstudium an der LMU München. **2011** Promotion an der LMU München. **2018** Habilitation im Fach Biostatistik an der Universität Erlangen-Nürnberg. Seit **2016** Wissenschaftler und Assessor und seit **2019** Leiter des Fachgebiets Biostatistik am Paul-Ehrlich-Institut, Langen.

Christel Kamp
1994–1999 Physikstudium an der Universität Münster. **2002** Promotion über die Dynamik komplexer biologischer Systeme an der Universität Kiel. **2019** Habilitation im Fach Bioinformatik an der Universität Frankfurt a. M.