Bioinformatik von Proteinsequenzen

von Dr. Horst Lohrer

Die Methoden der Molekularen Biologie, insbesondere die Sequenzierung von DNA und Proteinen, haben gigantische Mengen digitalisierter Daten generiert. Für die Analyse und Visualisierung dieser Daten stellt die Informatik die notwendigen Methoden. Die Aufgabe der Bioinformatik ist es, sicherzustellen, dass die biologischen Daten geordnet abgelegt werden, abrufbar sind und durch kreative Kombination zu neuen Erkenntnissen führen. Die vorliegende Unterrichtseinheit für den Biologieunterricht der gymnasialen Oberstufe bietet eine Einführung in die Thematik. Dabei liegt der Schwerpunkt auf dem Aspekt der Biologie in der Bioinformatik. Es wird dabei ausschließlich mit Proteinsequenzen gearbeitet. Die einzelnen Kapitel werden themenbezogen eingeleitet und erläutern notwendige Begriffe für die Benutzung der Algorithmen. Die praktische Arbeit mit öffentlich zugänglichen Datenbanken und Analyseprogrammen steht im Vordergrund der Arbeit. Den Beteiligten dieser Unterrichtseinheit wird ein erster Blick in die Methoden der Bioinformatik eröffnet, was hoffentlich zu eigenen Recherchen motiviert.

RAABE
KLASSE SCHULE

Impressum

RAABE UNTERRICHTS-MATERIALIEN Biologie Sek. II

Das Werk, einschließlich seiner Teile, ist urheberrechtlich geschützt. Es ist gemäß § 60b UrhG hergestellt und ausschließlich zur Veranschaulichung des Unterrichts und der Lehre an Bildungseinrichtungen bestimmt. Die Dr. Josef Raabe Verlags-GmbH erteilt Ihnen für das Werk das einfache, nicht übertragbare Recht zur Nutzung für den persönlichen Gebrauch gemäß vorgenannter Zweckbestimmung. Unter Einhaltung der Nutzungsbedingungen sind Sie berechtigt, das Werk zum persönlichen Gebrauch gemäß vorgenannter Zweckbestimmung in Klassensatzstärke zu vervielfältigen. Jede darüber hinausgehende Verwertung ist ohne Zustimmung des Verlages unzulässig und strafbar. Hinweis zu § 60b Abs. 3 UrhG: Das Werk oder Teile hiervon dürfen nicht ohne eine solche Einwilligung an Schulen oder in Unterrichts- und Lehrmedien (§ 60b Abs. 3 UrhG) vervielfältigt, insbesondere kopiert oder eingescannt, verbreitet oder in ein Netzwerk eingestellt oder sonst öffentlich zugänglich gemacht oder wiedergegeben werden. Dies gilt auch für Intranets von Schulen und sonstigen Bildungseinrichtungen. Die Aufführung abgedruckter musikalischer Werke ist ggf. GEMA-meldepflichtig.

Für jedes Material werden Fremdrechte recherchiert und ggf. angefragt.

Dr. Josef Raabe Verlags-GmbH
Ein Unternehmen der Klett-Gruppe
Rotebühlstraße 77
70178 Stuttgart
Telefon +49 711 62900-0
Fax +49 711 62900-60
mail@RAABE.de
www.raabe.de

Redaktion: Anne Zörlein
Satz: Rösel MEDIA GmbH & Co. KG, Karlsruhe
Bildnachweis Titel: Getty Images Plus/iStock/non-exclusive
Korrektorat: Josef Mayer

Bioinformatik von Proteinsequenzen

Niveau: weiterführend, vertiefend

von Dr. Horst Lohrer

Kompetenzprofil:

Kompeten	Anforderungs-bereich	Basiskonzept	Material
Fachwissen, Erkenntnisgewinnung, Kommunikation, Bewertung	I–III	Struktur und Funktion, Information und Kommunikation, Variabilität und Angepasstheit, Geschichte und Verwandtschaft	M 1–5

M 1 Evolution und Bioinformatik

Biologie ist das Fachgebiet für das Studium der Organismen und ihre Stammesgeschichte. Charles DARWIN erklärte 1859 in *The Origin of Species* das Prinzip der Evolution als die Lehre von der Veränderung von Populationen im Laufe der Generationen. Veränderungen entstehen durch spontane Mutationen der Erbsubstanz und werden an die Nachkommen weitergegeben. In der natürlichen Welt sind Ressourcen beschränkt und alle Individuen wetteifern im alltäglichen Überlebenskampf (*struggle for survival*) um ihre Lebensgrundlage. Individuen mit zufällig günstigen Eigenschaften werden erfolgreicher sein als andere Individuen ihrer Population und dessen relativ mehr Nachkommen (*fitness*) in die Population entlassen. Nach vielen Generationen wird nahezu die gesamte Population die vorteilhaften Eigenschaften besitzen. Erfolgreiche Populationen einer Art werden sich in diesem Wettlauf der natürlichen Selektion immer weiter an ihre Umweltbedingungen anpassen und, reproduktive Isolation vorausgesetzt, neue Arten bilden.

Organismen werden in Arten geordnet. Durch die Evolution entsteht eine Verwandtschaft zwischen den Arten. Nahe Verwandtschaft von Arten zeigt sich in einem ähnlichen Erscheinungsbild, weil der letzte gemeinsame Vorfahre in der nahen Vergangenheit (wenige Mio. Jahre) lebte. Liegt der letzte gemeinsame Vorfahre weit in der Vergangenheit (viele Mio. Jahre), dann verschwimmen die Ähnlichkeiten durch Mutation und Anpassung und eine Verwandtschaft ist nicht mehr offensichtlich. Die klassische Evolutionsforschung hat durch den Nachweis von Ähnlichkeiten in homologen, anatomischen Strukturen die Verwandtschaft zwischen Arten gezeigt und daraus Stammbäume abgeleitet.

Seit den 1950er-Jahren liefert die Molekulare Biologie eine gigantische Menge an molekularen Daten, darunter vor allem Sequenzdaten (DNA und Proteine) eine wahre Explosion erlebt haben. Sequenzdaten bestehen aus einer Abfolge von Symbolen und sind deshalb digitalisierbar. Damit sind sie auch das richtige Material für digitale, computerbasierte Methoden der Speicherung, Analyse und Aufbereitung. Homologie der molekularen Strukturen erlaubt Stammbäume allein mithilfe mathematischer Algorithmen auszuarbeiten.

M 1a Hypothesen zur Entstehung von Aminosäure-Ketten

Über den Beginn der molekularen Evolution vor etwa 4 Mrd. Jahren gibt es nur Modelle. Eines davon, die Bildung autokatalytischer chemischer Reaktionszyklen[1], hat viele der folgenden Modelle beeinflusst. Ein chemischer Kreisprozess erhält sich bei Zufuhr von Energie selbst. Die einzelnen Reaktionsschritte laufen unter Katalyse ab. In den ersten 500 Mio. Jahren der Erdgeschichte sind chemische Nicht-Gleichgewichte entstanden, die die Voraussetzungen für die Entstehung komplexerer Moleküle bei Abwesenheit von freiem Sauerstoff bildeten. Membranartige Strukturen erlaubten die Aufrechterhaltung der spezifischen Bedingungen innerhalb der „Zelle". Wenn diese Systeme die Fähigkeit zur Replikation erlangten, führte die natürliche Selektion zur Vermehrung der effektivsten Systeme und letztlich zu einem einfachen „Einzeller"[2].

Nach diesem Grundmodell könnten „Proteine" in fortgeschrittenen Hyperzyklen folgendermaßen entstanden sein: Kurze, spontan gebildete *mini*-Peptide bildeten „Katalysatoren" des Reaktionszyklus und wurden auf unbekannte Weise für das System fixiert. Nicht jedes spontane *mini*-Peptid wird zur Katalyse fähig gewesen sein. Nehmen wir an, ein *tri*-Peptid kann einen bestimmten Reaktionsschritt katalysieren, dann kann die zufällige Verknüpfung von Aminosäuren dieses *tri*-Peptid liefern: Mit 20 verschiedenen Aminosäuren können 20^3 verschiedene *tri*-Peptide gebildet werden. Aus diesen 8.000 Varianten kann das System das „richtige" *tri*-Peptid „herausuchen". Selbst bei einem *penta*-Peptid kann das System aus 20^5 Varianten das „richtige" Peptid selektionieren. Aus diesen einfachsten Katalysatoren könnten durch stufenweise Addition zufälliger Aminosäuren unter gleichzeitiger Selektion die heutigen Biokatalysatoren, bestehend aus Ketten von Hunderten Aminosäuren, entstanden sein. Die stufenweise Entwicklung bei gleichzeitiger Selektion ist die heute allgemein akzeptierte Hypothese (H) der Evolution.

Die alternative Hypothese (Null-Hypothese, H_0) beruht auf dem Modell von Versuch und Irrtum. Danach hätten sich die funktionellen, langkettigen Proteine spontan und in einem Schritt gebildet. Dies würde bedeuten, dass für eine Kettenlänge von 100 Aminosäuren 20^{100} verschiedene Varianten (oder 2 x 10^{101}) in einem spontanen Prozess entste-

---

1 EIGEN und SCHUSTER, 1979

2 EIGEN und SCHUSTER, 1979

hen, aus denen die funktionelle Variante herausgesucht würde. Unser bekanntes We[illegible]l besteht aus etwa 10^{80} Protonen. Die gesamten Kernbausteine der Atome des Uni[illegible]rsums würden nicht ausreichen, die möglichen Peptid-Varianten herzustellen und sie im Test auf Katalyse einzusetzen. Dadurch ist die Null-Hypothese als Erklärung der Evolutionsp[illegible]esse sehr unwahrscheinlich. Sie wird in der Bioinformatik als Alternative für st[illegible]tistische Berec[illegible]nungen eingesetzt, um Analysen auf der Basis der stufenweisen Evolution (H[illegible]) zu bewerten.

M 1b Information

Ketten aus Aminosäuren sind lineare Folgen von Symbolen und besitzen deshalb einen Informationsgehalt. Ihr Informationsgehalt kann beschrieben werden durch die durchschnittliche Zahl von Ja-/Nein-Entscheidungen, die für ihre Identifizierung notwendig ist. Die mathematische Beschreibung des Informationsgehaltes (I) einer Aminosäure-Sequenz ist abhängig von der Länge der Sequenz und der Wahrscheinlichkeit des Auftretens jedes Bausteins (p_i, ausgedrückt als Logarithmus zur Basis 2, wodurch die Einheit *binary digit* (= *bit*) erhalten wird). Es gilt:

$$I = \sum_{i=1}^{n} p_i \log_2 p_i$$

N = Kettenlänge

p_i = Wahrscheinlichkeit für einen bestimmten Baustein

Beispiel: für ein Peptid aus 30 Aminosäuren (Wahrscheinlichkeit des Einbaus einer bestimmten Aminosäure bei 20 möglichen Aminosäuren $p_i = 1/20$):

$$I = -30 \times \log_2 \frac{1}{20} = -30 \log_2 (0{,}05) = (-30) \times (-4{,}32) = 129{,}6 \textit{ bits}$$

Der Fluss des Informationsgehaltes von DNA zu mRNA und zu Proteinen lässt sich mit der obigen Gleichung sehr gut verfolgen. Für das menschliche Verständnis hat Information jedoch auch immer eine Bedeutung. Die Bedeutung einer Information eröffnet sich dem Empfänger nur in einem passenden Kontext. Die Bedeutung einer Information in einer für den Empfänger fremden Sprache kann ihre Bedeutung nicht entfalten. Für die Extraktion von Bedeutung aus einer Abfolge von Symbolen gibt es in der Molekularen Biologie keine begleitende Theorie.

M 1c Fragen zu M 1

1. Vier Bausteine mit den Symbolen „1“, „2“, „3“ und „4“ sollen zu einem Tripl[illegible]tt (drei Bausteine) verknüpft werden. Wie viele Varianten gibt es? **Berechnen** Sie, wie groß die Wahrscheinlichkeit für die Bildung folgender Varianten ist, wenn der Einbau von jedem Baustein gleich wahrscheinlich ist: Tripletts: „1-1-1“; „2-2-2“ und „2-3[illegible]

2. **Erklären** Sie, wie sich die Wahrscheinlichkeit für die Bildung des Tripletts „2-3-4“ ändert, wenn der Einbau des Bausteins „4“ doppelt so wahrscheinlich ist wie der Einbau von „2“ oder „3“.

3. Im Hypothalamus der Wirbeltiere werden von neurosekretorischen Zellen verschiedene Peptid-Hormone (*releasing hormone*, RH) gebildet, zur Hypophyse weitergeleitet und in den Blutstrom abgegeben. Die RH binden an Rezeptoren auf den Zielzellen und aktivieren über Signalketten ein genetisches Programm zur Herstellung spezifischer Hormone. Gonadoliberine (GRH) zum Beispiel stimulieren die Freisetzung der Gonadotropine LH (Luteotropes Hormon) und [illegible] (Follikelstimulierendes Hormon). Die Aminosäure-Sequenz des GRHs des Huhns lautet:

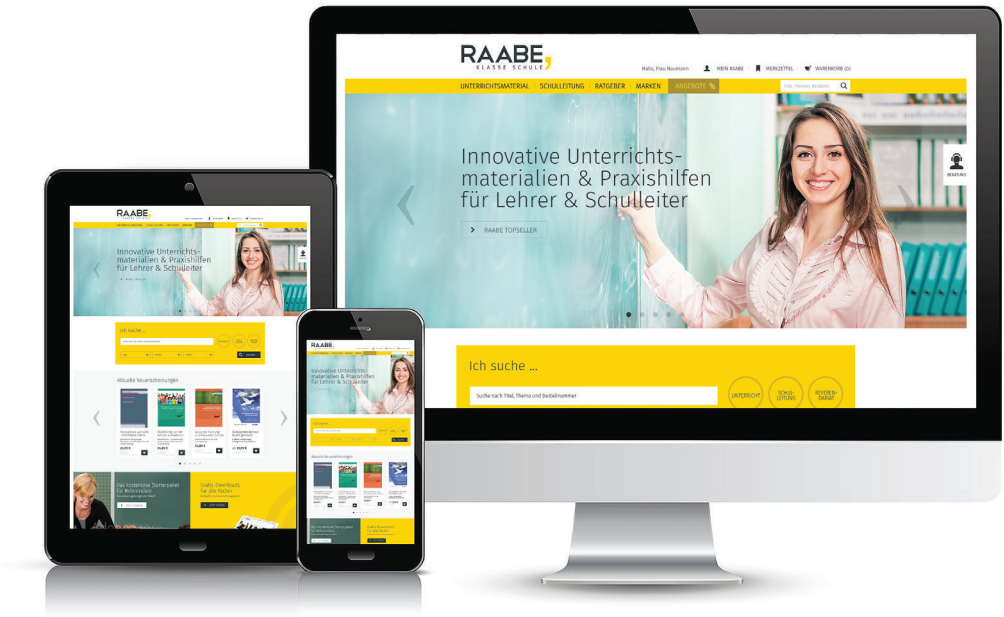
   Glu-His-Trp-Ser-Tyr-Gly-Leu-Gln-Pro-Gly.

   Physiologische Dosen des Hormons führen im Huhn zu 100 % Freisetzung von LH, im Schaf jedoch nur zu 3 %[3]. **Berechnen** Sie den Informationsgehalt des GRHs des Huhns und erläutern Sie die unterschiedliche Freisetzung von LH in Huhn und Schaf.

[3] PENZLIN 2014

RAABE
KLASSE SCHULE

Der RAABE Webshop: Schnell, übersichtlich, sicher!



Wir bieten Ihnen:

- Schnelle und intuitive Produktsuche
- Übersichtliches Kundenkonto
- Komfortable Nutzung über Computer, Tablet und Smartphone
- Höhere Sicherheit durch SSL-Verschlüsselung

Mehr unter: www.raabe.de