Metadaten

MIxS: neue Standards für Sequenz- und Umweltdaten

FRANK OLIVER GLÖCKNER, RENZO KOTTMANN
MAX-PLANCK-INSTITUT FÜR MARINE MIKROBIOLOGIE BREMEN UND JACOBS-UNIVERSITÄT BREMEN

Standen vor einigen Jahren noch die Entschlüsselung einzelner Gene oder Genome durch wenige Spezialisten im Vordergrund, sind es heute die überall verfügbaren, enormen Sequenzierkapazitäten, die die Lebenswissenschaften prägen. Um diese Flut an Daten in biologisches Wissen zu überführen, müssen diese mit Zusatzinformationen (Metadaten) verknüpft werden. Entsprechende Standards für Genome, Metagenome und Markergene zu entwickeln und zur Anwendung zu bringen, hat sich das Genomic Standards Consortium (GSC) zur Aufgabe gemacht [1].

In den öffentlichen Datenbanken sind bereits heute Tausende Genome, Hunderte Metagenome und unzählige Sequenzen von Markergenen zu finden. Die Geschwindigkeit, mit der neue Sequenzdaten dazukommen, ist atemberaubend. Gerade die *Next Generation*-Sequenzierungsmethoden erlauben es, in kürzester Zeit Datenmengen zu erzeugen, die vor ein paar Jahren noch undenkbar waren. Optimisten sagen voraus, dass in nur wenigen Jahren die DNA praktisch aller Lebewesen auf der Erde entziffert ist. Dabei spiegelt jeder Datensatz die speziellen Eigenschaften eines Organismus oder einer Organismengemeinschaft wider. Jeder dieser Datensätze ist somit ein weiterer Baustein, um den „Katalog des Lebens" unserer Erde zu erstellen. Die Sequenzinformation alleine reicht jedoch nicht aus, um die einzigartigen Entstehungsgeschichten, Lebensräume und biologischen Eigenschaften verstehen zu lernen. Für einen nachhaltigen Erkenntnisgewinn sind vor allem Zusatzdaten wie der Ort und Zeitpunkt der Probenahme sowie grundlegende Experimental- und Umweltparameter, die Metadaten, entscheidend. Dazu reicht es jedoch nicht, irgendwelche Daten zu erheben und irgendwo abzulegen. Es müssen Richtlinien entworfen werden, welche Metadaten für eine Probe oder Studie entscheidend sind und wie diese strukturiert und damit zugreifbar in öffentlich zugänglichen Datenbanken abgelegt werden können.

Bereits 2005 fand sich eine Gruppe von internationalen Wissenschaftlern in dem für jeden offenen GSC zusammen. Das Ziel des GSC ist es, Richtlinien für eine möglichst kompakte, aber dennoch repräsentative Menge an wünschenswerten Zusatzdaten für Sequenzinformationen zu entwerfen. Daraus entstanden zunächst der MIGS(*minimum information about a genome sequence*)- und MIMS(*minimum information about a metagenome sequence*)-Standard für Genom- und Metagenominformationen [2]. Nach zwei weiteren Entwicklungsjahren konnte das Konsortium jetzt den MIMARKS(*minimum information about a marker gene sequence*)-Standard und die MIxS-Spezifikationen (*minimum information about any (x) sequence*) veröffentlichen [3]. Dabei ist MIxS die übergeordnete Spezifikation der bisherigen drei Standards (MIGS/MIMS/MIMARKS). Die Idee ist dabei, dass man je nach biologischem

MIxS-Checklisten

	MIGS					MIMS	MIMARKS	
Item	EU	BA	PL	VI	ORG	ME	SU	SP
Project name	M	M	M	M	M	M	M	M
Collection date	M	M	M	M	M	M	M	M
Geographic location	M	M	M	M	M	M	M	M
Environment	M	M	M	M	M	M	M	M
Ploidy	M	-	-	-	-	-	-	-
Estimated size	M	X	X	X	X	-	-	-
Target gene	-	-	-	-	-	-	M	M

Umweltpakete

Item	Human-gut	Soil	Water
Age	X	-	-
Body site	X	-	-
Elevation	-	M	C
Depth	-	M	M
Horizon	-	X	-
pH	-	X	X
Salinity	-	-	X

MIMARKS-Survey (Umweltstudie) Soil-Checkliste

Item	SU
Project name	M
Collection date	M
Geographic location	M
Environment	M
Target gene	M
Elevation	M
Depth	M
Horizon	X
pH	X

◀ **Abb. 1:** Übersicht über die MIxS-Checklisten (*minimum information about any (x) sequence*, braun) des Genomic Standards Consortium und die Möglichkeit, diese mit Umweltpaketen zu kombinieren (blau). EU, Eukarya; BA, Bacteria/Archaea; PL, Plasmid; VI, Virus; ORG, Organelle; ME, Metagenome; SU, Survey (Umweltstudie); SP, Spezies. M, *mandatory* (notwendig); C, *conditional mandatory* (bedingt notwendig); X, optional; ein Strich bedeutet nicht anwendbar. Als Beispiel wurde die MIMARKS-Survey-Checkliste ausgewählt und mit dem Umweltpaket „Boden" kombiniert.

DOI: 10.1007/s12268-011-0115-2

Ursprung der Sequenzen einen entsprechenden Standard wählt, etwa für Eukaryoten, Bakterien oder Viren (**Abb. 1**). Daraus ergibt sich, welche Metadaten auf jeden Fall (M) erfasst und abgelegt werden sollten und welche empfohlen (X) werden. Zusätzlich wurden 14 Umweltpakete zur Beschreibung bestimmter Habitate wie Boden, Wasser oder den Menschen eingeführt.

Die Kombination aus allgemeinen Metadatenspezifikationen und spezifisch auf die Studie zugeschnittenen Zusatzdaten machen die MIxS-Standards zu einem leistungsfähigen und flexiblen Instrument für die Metadatenerhebung. Bereits heute erkennen die öffentlichen Sequenzdatenbanken INSDC (International Nucleotide Sequence Database Collaboration [EBI-ENA, GenBank und DDBJ]) die MIxS-Standards an, und jeder kann Metadaten strukturiert in den INSDC ablegen.

Mit der Veröffentlichung der MIxS-Standards steigt die Hoffnung der Forscher, in Zukunft nicht nur die schiere Quantität der Sequenzdaten, sondern auch deren Qualität durch entsprechende Metadaten nachhaltig zu verbessern.

Literatur

[1] Yilmaz P, Gilbert JA, Knight R et al. (2011) The genomic standards consortium: bringing standards to life for microbial ecology. Isme J. http://dx.doi.org/10.1038/ismej.2011.39
[2] Field D, Garrity G, Gray T et al. (2008) The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26:541–547
[3] Yilmaz P, Kottmann R, Field D et al. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29:415–420


Frank Oliver Glöckner (links) und Renzo Kottmann

Korrespondenzadresse:
Prof. Dr. Frank Oliver Glöckner
Max-Planck-Institut für Marine Mikrobiologie
Celsiusstraße 1
D-28359 Bremen
Tel.: 0421-2028-970
Fax: 0421-2028-580
fog@mpi-bremen.de
www.microbial-genomics.de
http://gensc.org