Integration biologischer Sequenzdaten zum Aufbau eines Genomwarehouses

André Bergholz Johann Christoph Freytag

Daten in der Molekularbiologie werden in sehr heterogener Form gesammelt. Zum einen werden unterschiedliche Bereiche, wie Sequenzen, 3D-Strukturen und Experimente abgedeckt. Aber auch innerhalb eines Bereiches, wie dem der DNA-Sequenzen, werden sehr unterschiedliche Daten, von der eigentlichen Sequenz über Eigenschaftenprofile bis zum Literaturverweis, gesammelt. Eine weitere Art der Heterogenität entsteht dadurch, daß nicht einheitlich geregelt wird, was als ein Sequenzeintrag angesehen wird. Es werden DNA und RNA vermischt, Exons und Introns tauchen nochmals als Teil eines kompletten Gens auf.

In unserem Projekt wollen wir Genom- und Proteomsequenzdaten unter einem einheitlichen Modell speichern und zur Weiterverarbeitung bereitstellen. Dabei wollen wir, im Gegensatz zu Black-Box-Systemen wie PC/Gene oder HUSAR, eine offene Umgebung mit Schnittstellen zur Entwicklung eigener Applikationen schaffen. Es soll eine homogene, logisch einheitliche Form der Daten entstehen. Relationale Datenbanktechnologie soll helfen, den Zugriff auf die Daten einerseits deklarativ, andererseits effizient zu gestalten. So können einerseits ausdrucksstärkere Anfragen an Daten gestellt werden, andererseits werden intelligente Indizierungstechniken ermöglicht.

Als Grundlage orientieren wir uns an den Datensammlungen EMBL und SWISS-PROT, versuchen aber schrittweise, eher die reale Welt als Datensammlungen zu modellieren. Zum Modellieren verwenden wir verwenden die Entity-Relationship-Technik. Die zentrale Entität, vergleichbar mit einer Faktentabelle, ist in unserem Modell die Entität „Lineares Polymer“, die einen Oberbegriff für DNA, RNA und Protein darstellt. Um diese Entität scharen sich die Dimensionen: Spezies mit Klassifikation, Eigenschaftenprofil, Literaturverweise und Schlüsselwörter. Ein lineares Polymer ist mir verschiedenen elementaren Attributen, wie Beschreibung oder Datum der letzten Änderung versehen. Es ist einer Spezies zugeordnet, und Spezies sind Speziesklassen zugeordnet. Die Speziesklassen sind hierarchisch angeordnet. Eigenschaften stehen in Beziehung zu linearen Polymeren. Diese Beziehung ist mit Attributen, wie dem Ort der Eigenschaft oder einem bestimmten Wert einer Eigenschaft versehen. Literaturreferenzen verfügen über eine Vielzahl von optionalen Attributen, um die Heterogenität zwischen den verschiedenen Arten der Literaturreferenzen, wie Artikel, Patente oder Dissertationen zu behandeln. Die Schlüsselwortverwaltung ordnet Schlüsselwörter einem Referenzschlüsselwort zu und stellt dieses in Beziehung zu linearen Polymeren. Abbildung 1 zeigt unseren Entity-Relationship-Ansatz (ohne Attribute). Gegenwärtig arbeiten wir an einer Verbesserung des Modells in zwei Richtungen: Erstens wollen wir das Modell näher an die reale Welt und weiter weg von spezifischen Datensammlungen bringen. Dies soll durch explizite Modellierung von Entitäten wie Allel, Spliceform oder Proteinfamilie erreicht werden. Zweitens soll jede Information zu einer Informationsquelle in Beziehung gestellt werden, so daß auch widersprüchliche Informationen in der Datenbank verwaltet werden können. Informationsquellen sollen dann sowohl Datensammlungen als auch Literaturreferenzen umfassen.

Unsere auf diesem Modell basierende Datenbank "Biological Sequences Integrated" (BSI) wurde im System Sybase implementiert. Die Abbildung 2 verdeutlicht diesen Prozeß. Die Datenbank wurde mit Daten der Sammlungen EMBL und SWISS-PROT gefüllt und hat einen Umfang von ca. zwei Gigabyte. Die Datenbank bildet die Grundlage für ein erstes Anwendungsprogramm: das auf der Embedded-SQL Schnittstelle basierendes System "No Name Sequence Analysis Tools" (NNSAT). Neben der Anbindung an die BSI Datenbank werden gängige Dateiformate wie EMBL unterstützt. NNSAT bietet elementare Funtionalität für Sequenzen, wie z. B. Translation oder Berechnung des Gegenstranges einer DNA Sequenz. Außerdem wird Alignierung nach Needleman und Wunsch in zwei Varianten und mit verschiedenen Alignmentmatrizen bereitgestellt.

Gegenwaertig konzentriert sich unsere Arbeit auf zwei Teilprojekte: Mit der Portierung der Datenbank auf das System DB2 sollen die Vorteile von Parallelität und objektrelationaler Technologie nutzbar gemacht werden. In einer Diplomarbeit werden zudem Indizierungstechniken, die einen effizienteren Zugriff auf DNA-Sequenzen ermöglichen sollen, untersucht.

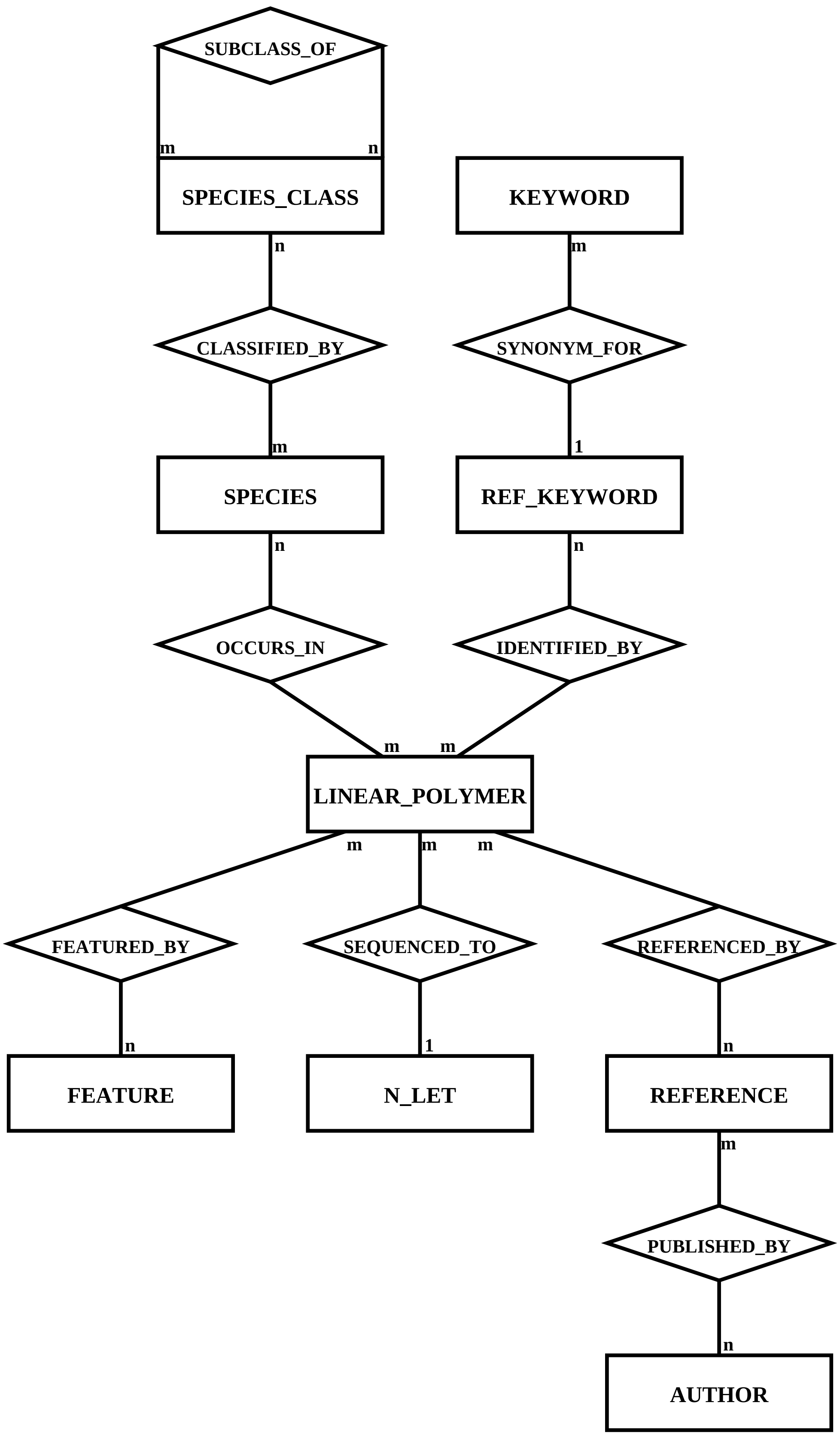

Abbildung 1: Der Entity-Relationship Ansatz für biologische Sequenzdaten

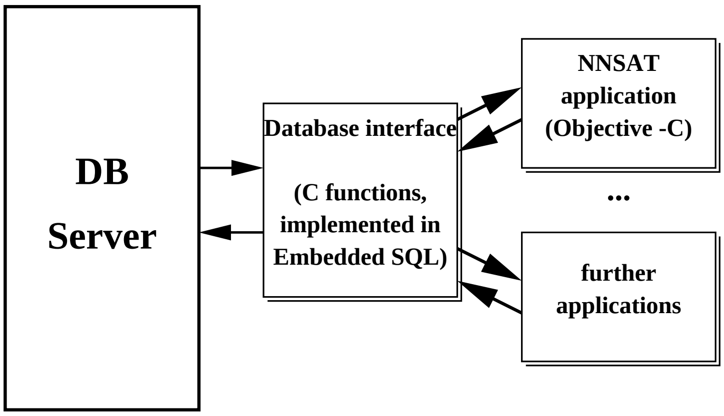

Abbildung 2: Der Aufbau der Datenbank BSI