Eclipse4Bio

Projektgruppe des Lehrstuhls für Programmiersysteme (LS5)

1. Thema

Eclipse-basierte Open-Source-Plattform für Bioinformatik-Prozesse

2. Zeitraum

Sommersemester 2011 und Wintersemester 2011/2012

3. Umfang

Jeweils 8 Semester-Wochenstunden

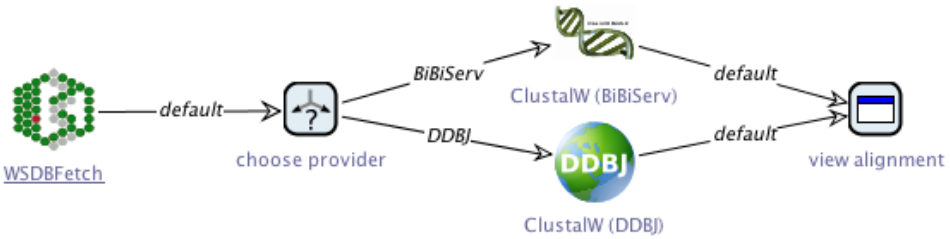
4. Veranstalter

Anna-Lena Lamprecht, OH14, Raum 131, Tel. 7736
Ralf Nagel, OH14, Raum 135, Tel. 5806
Sven Jörges, OH14, Raum 130, Tel. 7757

5. Aufgabe

In der biologischen Forschung, insbesondere in den Bereichen der Genetik und Molekularbiologie, fallen heute Datenmengen an, die ohne Computerunterstützung nicht zu bewältigen sind. Aus dieser Notwendigkeit heraus hat sich die Disziplin der *Bioinformatik* entwickelt, die sich weiter in *theoretische* und *angewandte* Bioinformatik einteilen lässt [1]. Während sich die theoretische Bioinformatik hauptsächlich mit der Entwicklung von Algorithmen für bestimmte biologische Probleme befasst, geht es in der angewandten Bioinformatik darum, diese Verfahren zu implementieren und für die Forschergemeinde in großem Rahmen nutzbar zu machen.

Als ein Resultat der Bestrebungen in der angewandten Bioinformatik stehen heute viele *Web Services* zur Verfügung, die freien Zugriff auf biologische Datenbanken oder Bioinformatik-Analysealgorithmen anbieten. Die in Forschungsprojekten anfallenden genetischen und molekularbiologischen Daten werden häufig mit Hilfe mehrerer solcher öffentlich verfügbaren Dienste analysiert. Eine spezielle Kombination solcher Schritte zur Lösung eines bestimmten Analyseproblems definiert einen *Workflow*. Häufig werden diese Workflows von den Forschern manuell ausgeführt, was jedoch zeitaufwändig und fehleranfällig ist. Gerade bei der wiederholten Anwendung auf zahlreiche Datensätze ist eine Automatisierung der Workflowausführung erstrebenswert.



1: Beispiel-Workflow.

Abbildung 1 zeigt einen Beispiel-Workflow, bei dem zunächst DNA-Sequenzen aus einer Datenbank geladen werden (WSDBFetch), dann wahlweise mit dem ClustalW-Algorithmus auf dem BiBiServ (Bielefeld University Bioinformatics Server) oder beim DDBJ (DNA Data Bank of Japan) ein Sequenzalignment berechnet und schließlich angezeigt wird.

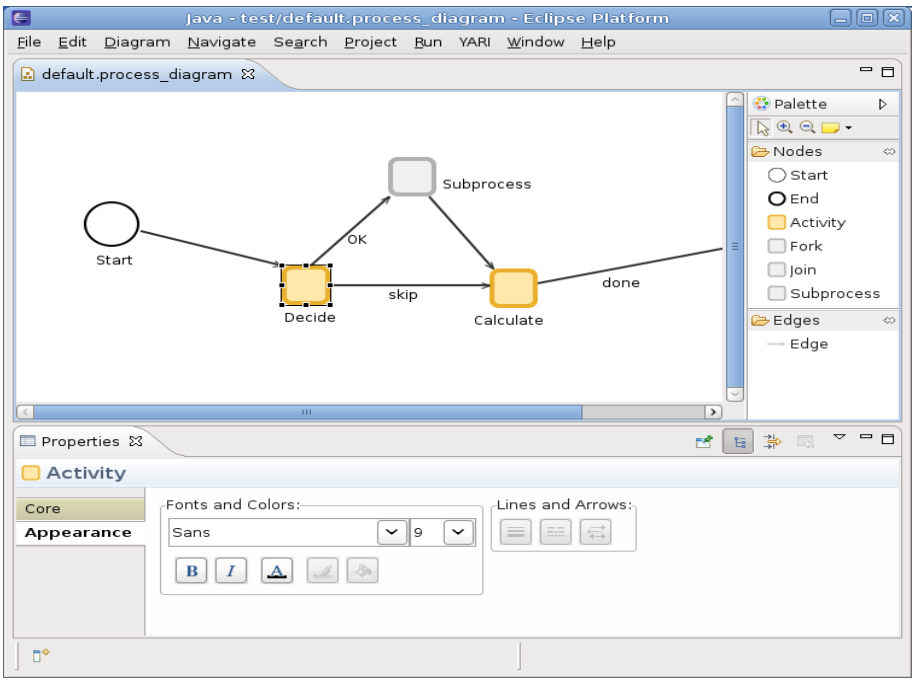
In den letzten Jahren wurden daher in der Bioinformatik verschiedene Systeme entwickelt, die die Entwicklung von Bioinformatik-Workflows gezielt unterstützen (siehe z.B. [2]). Sie sind zumeist an konkreten Bedürfnissen orientiert entstanden und setzen nicht auf Geschäftsprozess-Standards (siehe z.B. [4]) auf. Mit **Eclipse4Bio** soll eine auf offenen Standards basierende Plattform für Bioinformatik-Workflows realisiert und evaluiert werden. Sie soll dazu mit bestehenden Systemen umfassend verglichen werden, wobei die Frage im Vordergrund steht, inwiefern sich speziell für Geschäftsprozesse entwickelte Technologien auch für wissenschaftliche Workflows einsetzen lassen.

Technischer Hintergrund

Eclipse [5] ist nicht nur eine Java-IDE, sondern vielmehr eine Sammlung von hochentwickelten Software-Werkzeugen. In der PG Eclipse4Bio sollen vor allem Technologien aus dem Eclipse Modeling Project (EMP) [6] zum Einsatz kommen, die die Arbeit mit domänen-spezifischen Sprachen (domain-specific languages, DSLs) und modellgetriebener Entwicklung (model-driven development, MDD) unterstützen. Dazu gehören beispielsweise grafische Editoren wie der *BPMN Modeler* [7] oder das *Java Workflow Tooling* (JWT, [8]), die elementare Funktionalitäten zur grafischen Prozessmodellierung zur Verfügung stellen. Sie werden ergänzt von Übersetzungsmechanismen in Standard-Prozessbeschreibungsnotationen wie BPMN [4], die in entsprechenden Ausführungsumgebungen direkt lauffähig sind. Insbesondere bietet das EMP zahlreiche Schnittstellen zur Erweiterung an, sodass das Framework gezielt an ein bestimmtes Anwendungsgebiet angepasst werden kann. Die PG Eclipse4Bio soll diese Technologien einsetzen, um an die Bedürfnisse der Worfklow-Entwicklung für Bioinformatik zugeschnittene Editoren zu entwickeln.

Die Workflows sollen verfügbare Bioinformatik-Dienste ansteuern und zusammen mit selbst entwickelten Analysemethoden die Werkzeugpalette des Eclipse4Bio-Editors bilden. Ein zentraler Bestandteil der PG-Arbeit ist also die automatisierte Integration von öffentlich verfügbaren Bioinformatik-Diensten, wie sie von großen Institutionen wie dem European Bioinformatics Institute (EBI), aber auch von kleineren Einrichtungen auf der ganzen Welt angeboten werden. Dabei ist es zunächst ausreichend, die *BioCatalogue*-Dienste zu integrieren. Der BioCatalogue [3] ist mit über 1700 registrierten Web Services der derzeit umfangreichste Verzeichnisdienst der Bioinformatik. Über eine API kann das Verzeichnis bequem durchsucht und die Services angesprochen werden.

Die in Eclipse4Bio integrierten Dienste sollen eine Palette von (grafisch repräsentierten) Komponenten bilden, aus denen Anwender vielfältige Bioinformatik-Workflows per *drag-and-drop* zusammenstellen können. Mit zusätzlichen Kontrollfluss-Bausteinen sind auch bedingte Anweisungen oder Schleifen darstellbar. Ein solcher domänenspezifischer Editor kann z.B. mit Hilfe des Eclipse Graphical Modeling Framworks (GMF) [9] erzeugt werden, das ebenfalls zum EMP gehört. Abbildung 2 zeigt, wie ein einfacher mit GMF erzeugter Editor aussehen könnte: Auf der großen „Zeichenfläche“ findet die Workflow-entwicklung statt. Rechts davon ist die Palette der verfügbaren Komponenten dargestellt. In der Properties-Ansicht an der unteren Seite des Fensters können Konfigurationen vorgenommen werden.



2: Mit GMF erzeugter Editor.

Ein weiteres Ziel der PG ist es, die im grafischen Editor erstellten Workflows auch direkt ausführbar zu machen. Dies kann zum einen als Java-Prozess direkt in Eclipse oder durch eine spezialisierte Prozess-Engine (wie z.B. eine BPEL-Engine) geschehen. Die Ausführung als Java-Prozess kann leicht beobachtet und debugged werden. Execution Engines eignen sich dagegen für lang dauernde Prozesse oder die Serienausführung von vielen Prozessen.

Insgesamt soll mit Eclipse4Bio durch Auswahl geeigneter Eclipse-Technologien eine nahtlose Integration von grafischem Editor, angepasster Tool-Palette und Ausführungsumgebung erreicht werden, sodass Bioinformatik-Analyseprozesse auch ohne besondere Programmierkenntnisse modelliert und ausgeführt werden können. Damit sollen dann solche Workflows umgesetzt werden können, wie sie z.B. in myExperiment [10] verzeichnet sind.

Darüber hinaus ist die Einbindung semantikorientierter Technologien denkbar: Im BioCatalogue werden Dienste nicht nur gelistet, sondern auch annotiert, d.h. mit *Metadaten* versehen, die die Services mit Hilfe eines kontrollierten Vokabulars beschreiben. Diese semantischen Informationen können beispielsweise für komplexe Discovery-Mechanismen genutzt werden, oder die Grundlage für die Anwendung von Synthese- und Planungstechnologien für die (semi-) automatische Workflowentwicklung bilden [11].

6. Teilnahmevoraussetzungen

- Fundierte Kenntnisse in mindestens einer objektorientierten Programmiersprache, z. B. Java **(V)**
- Kenntnisse in dem Gebiet Software-Design/Implementierung durch erfolgreiche Teilnahme an mindestens einer der Vorlesungen “Formale Methoden des Systementwurfs”, “Darstellung, Verarbeitung und Erwerb von Wissen”, “Effiziente Algorithmen”, “Modellgestützte Analyse und Optimierung”, “Mensch-Maschine-Interaktion”. **(V)**

- Kenntnisse in Eclipse-Werkzeugen und -Programmierung **(W)**
- Kenntnisse über Geschäftsprozesse, Web Services, Semantic Web **(W)**

Legende

(V) Voraussetzung
(W) Wünschenswert

Für die Teilnahme an der PG sind keine Vorkenntnisse in Biologie/Bioinformatik notwendig!

7. Minimalziel

Im Rahmen der PG sollen mindestens die folgenden Punkte umgesetzt werden:

- Entwicklung eines Eclipse-basierten, erweiterbaren, grafischen Editors für Bioinformatik-Workflows
- automatisierte Integration der BioCatalogue-Dienste in den Editor (Verwendung einfacher Codegeneratoren wie z.B. WSDL2Java)
- Ausführung von Beispielprozessen
  - direkt als Java-Code
  - durch Export in eine Prozesssprache (wie z.B. BPEL)
- Vergleich von Geschäftsprozess-Technologien mit wissenschaftlichen Worfklow-Systemen

8. Literatur

[1] P. Selzer, R. Marhöfer, A. Rohwer: *Angewandte Bioinformatik: Eine Einführung (Springer-Lehrbuch)*. Springer, Berlin; Auflage: 1 (17. September 2003)

[2] A.-L. Lamprecht, T. Margaria, B. Steffen: *Bioinformatics: Processes and Workflows.* Encyclopedia of Software Engineering, Auerbach Publications, 2010.

[3] The BioCatalogue: a curated catalogue of Life Science Web Services http://www.biocatalogue.org/

[4] T. Allweyer: *BPMN 2.0 - Business Process Model and Notation: Einführung in den Standard für die Geschäftsprozessmodellierung.* Books on Demand; Auflage: 2. Auflage. (3. November 2009)

[5] Eclipse: http://www.eclipse.org/

[6] R. C. Gronback: Eclipse Modeling Project - A Domain-Specific Language (DSL) Toolkit. Addison-Wesley Longman, Amsterdam; Auflage: 1 (6. März 2009)

[7] Eclipse BPMN Modeler; http://www.eclipse.org/bpmn/

[8] Eclipse Java Workflow Tooling (JWT), http://www.eclipse.org/jwt/

[9] Eclipse Graphical Modeling Framework (GMF), http://www.eclipse.org/modeling/gmp/

[10] myExperiment Virtual Research Environment - http://www.myexperiment.org/

[11] A.-L. Lamprecht, S. Naujokat, T. Margaria, B. Steffen: Semantics-based composition of EMBOSS services. BMC Bioinformatics, 2010

9. Rechtlicher Hinweis

Die Ergebnisse der Projektarbeit und die dabei erstellte Software sollen der Fakultät für Informatik uneingeschränkt für Lehr- und Forschungszwecke zur freien Verfügung stehen. Darüber hinaus sind keine Einschränkungen der Verwertungsrechte an den Ergebnissen der Projektgruppe und keine Vertraulichkeitsvereinbarungen vorgesehen.