Einsatz von FlowMark™ in der Molekularbiologie

Angelo Brayner, Mathias Weske*
Institut für Wirtschaftsinformatik
Universität Münster

Zusammenfassung

In diesem Papier untersuchen wir einen neuen Einsatzbereich für Workflow-Management Systeme, nämlich die Naturwissenschaften. Wir schildern unsere Erfahrungen bei der Modellierung eines naturwissenschaftlichen Workflows aus dem Bereich der Molekularbiologie unter Verwendung des Produkts FlowMark der Firma IBM und leiten daraus Eigenschaften für Workflow-Management Systeme ab, die für wissenschaftliche Anwendungen relevant sind.

1 Einführung

FlowMark[1] ist ein Workflow-Management System, das die Spezifikation strukturierter Arbeitsabläufe und ihre kontrollierte Ausführung ermöglicht [3]. Typisches Einsatzgebiet von FlowMark ist eine kommerzielle Büro-Umgebung, mit dem Ziel der Modellierung sowie der kontrollierten und effizienten Ausführung von Geschäftsprozessen.

In diesem Bericht wollen wir unsere Erfahrungen mit FlowMark in einem neuen Einsatzbereich von Workflow-Management Systemen, den Naturwissenschaften, schildern. Konkret geht es um einen prototypischen Einsatz von FlowMark[2] bei der Modellierung und kontrollierten Ausführung von Experimenten in der Molekularbiologie. Die zentrale Fragestellung lautet, inwiefern ist ein für den Bereich von Business-Anwendungen entwickeltes System geeignet, in den Naturwissenschaften auftretende Problemstellungen zu lösen, bzw., welche Veränderungen sind notwendig, um Workflow-Management Systeme auch in neuen Problemfeldern einsetzen zu können.

Bei der konkreten Anwendung handelt es sich um ein komplexes Experiment, das das Ziel verfolgt, ein DNA-Molekül zu sequenzieren, d.h. seine Basenfolge zu bestimmen. Experimente dieser Art bilden einen wichtigen Teil gentechnologischer Forschung, die etwa im Rahmen des Human Genome Projects durchgeführt wird [1, 7]. Da heutige labortechnische

*E-mail: {brayner,weske}@uni-muenster.de

[1] FlowMark ist eingetragenes Warenzeichen der Firma IBM.

[2] Installation: FlowMark (Version 2, Release 1.0), Pentium 90, 16MB RAM, OS/2, Version 3.

Beschränkungen lediglich die Sequenzierung von DNA-Molekülen der Länge einiger hundert Basen erlauben, wird die gesamte Sequenz zunächst in kleinere Teilsequenzen aufgespalten, um sie anschließend computerunterstützt (etwa durch das Programmsystem FAT [5]) zusammenzusetzen. Dieser Vorgang wird als *Fragment Assembly* bezeichnet; er bildet den Kern des in diesem Bericht untersuchten DNA-Sequencing Workflows. Dieser Workflow wurde durch 23 Aktivitäten modelliert, die sich auf unterschiedlichen Abstraktionsebenen befinden und manuell, semi-automatisch oder automatisch durch das FAT-System [5] von C++ Programmen unter OS/2 und DOS ausgeführt werden.

Dieser Bericht ist wie folgt strukturiert. Abschnitt 2 gibt eine kurze Einführung in Workflow-Management im Bereich der Naturwissenschaften und beschäftigt sich mit der Modellierung des wissenschaftlichen Workflows "DNA-Sequencing", die unabhängig von einer konkreten Implementation vorgenommen wurde. Abschnitt 3 beschreibt zunächst einige allgemeine Erfahrungen, die wir bei dem Einsatz von FlowMark machen konnten, bevor wir über die Implementation des DNA-Sequencing Workflows in FlowMark berichten. Dieses Papier schließt mit einer Zusammenfassung, wobei auch auf mögliche Erweiterungen von FlowMark hingewiesen wird, die wir anhand unserer konkreten Erfahrungen für wünschenswert erachten.

2 Workflow-Management in den Naturwissenschaften

2.1 Wissenschaftliche Workflows

Der Fortschritt in den Naturwissenschaften hängt in zunehmendem Maß von dem effizienten Einsatz von Computertechnologien ab. In vielen naturwissenschaftlichen Disziplinen werden derzeit Computer zur Steuerung von Versuchsapparaturen, zur Sammlung und Verwaltung von Datensätzen und zur Validierung und Analyse von Ergebnissen eingesetzt. Der Einsatz von Computern bezieht sich allerdings meist auf spezielle und isolierte Fragestellungen, d.h., einzelne Aktivitäten werden durch Computer und die entsprechenden Applikationen unterstützt.

Vor diesem Hintergrund fällt bei der Betrachtung wissenschaftlicher Arbeitsweisen und -methoden auf, daß sie einen starken Prozeß-Charakter aufweisen. Ein typischer Prozeß im Rahmen einer wissenschaftlichen Anwendung beginnt mit der Datensammlung durch Experimente, die manuell, semi-automatisch oder automatisch durchgeführt werden. Anschließend erfolgen eine Reihe von Kalibrierungen und Validierungen der Daten, bis es schließlich zu ihrer Analyse und Interpretation kommt.

Nun liegt der Ansatz nahe, diese Prozesse durch Workflow-Management Systeme zu unterstützen. Ein diesbezüglicher Ansatz wurde kürzlich vorgeschlagen [4]. Bei dem Einsatz von Workflow-Management Systemen in wissenschaftlichen Anwendungen liegt das Haupt-Augenmerk nicht auf der Steigerung des Durchsatzes oder der Verringerung der Kosten (wie dies bei dem klassischen Einsatzfeld dieser Systeme im Zusammenhang mit Geschäftsprozessen zu beobachten ist), sondern auf der Dokumentation von Experimenten und – damit verbunden – der Wiederholbarkeit ihrer Ergebnisse. Große Bedeutung kommt dabei auch

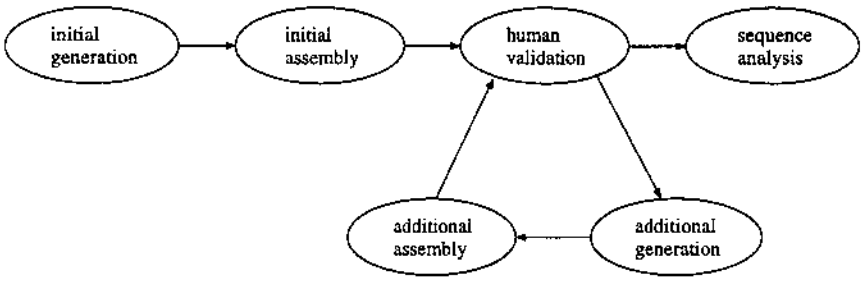

Abbildung 1: DNA Sequenzierung als Workflow.

der Wiederbenutzbarkeit von Prozeß-Beschreibungen zu. Weitere Eigenschaften von wissenschaftlichen Workflows und eine Referenz-Architektur *WASA* (Workflow-based Architecture to support Scientific Applications) wurden in [4] vorgestellt.

2.2 DNA Sequenzierung als Workflow

Nun beschreiben wir eine Beispiel-Anwendung eines wissenschaftlichen Workflows aus dem Bereich der Molekularbiologie, in welchem die Notwendigkeit der Computerunterstützung früh erkannt wurde und entsprechende Initiativen unterstützt werden [2, 8].

Wir werden im Rahmen dieses Berichts nicht auf die biologischen Grundlagen eingehen, sondern lediglich das generelle Verfahren und die Abbildung dessen in eine Workflow-Spezifikation vorstellen. Wir betrachten das bereits in der Einleitung kurz diskutierte Experiment "DNA Sequenzierung", das aus einer Reihe von Teil-Experimenten besteht und dessen Zielsetzung die Identifikation der Basensequenz eines DNA-Strangs ist [6].

Wir diskutieren nun die Abbildung dieses komplexen wissenschaftlichen Experiments als Workflow und die Unterstützung, die von einem Workflow-Management System in diesem Zusammenhang zu erwarten ist. Wie dies auch bei nicht-wissenschaftlichen Anwendungen zu beobachten ist, werden unterschiedliche Abstraktionsebenen der Modellierung ausgewählt; wir beginnen bei der obersten Ebene, die auch als *Toplevel-Workflow* bezeichnet wird (s. Abb. 1).

Anhand von Abbildung 1 beschreiben wir nun den generellen Ablauf des wissenschaftlichen Workflows. Zunächst werden aus dem zu untersuchenden DNA-Strang eine Menge von Fragmenten isoliert und diese sequenziert. Diese Aktivität wird als *initial generation* bezeichnet. Anschließend wird das FAT-Programmsystem [5] zur Assemblierung der Fragmente verwendet; dies erfolgt im Rahmen der Aktivität *initial assembly*. Anschließend erfolgt die semi-automatische Aktivität *human validation*, bei der eine Person anhand der Ausgaben des Assemblierungs-Schrittes entscheidet, ob dieser erfolgreich war oder ob zusätzliche Fragmente generiert (*additional generation*) und anschließend zu den bereits gewonnenen Teilsequenzen hinzugefügt werden müssen (*additional assembly*).

Die Verfeinerung von *human validation* ist in Abbildung 2 dargestellt. Bei ihr werden zunächst die Ergebnisse des vorhergehenden Assemblierungsschrittes betrachtet (*viewing results*), bevor die Grundlagen für die Assemblierung, die sogenannten *alignments*, überprüft

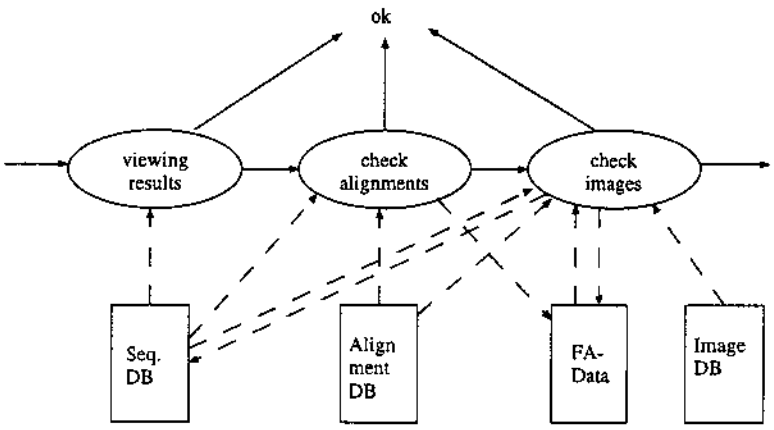

Abbildung 2: Human Validation. Unter Verwendung einer Sequenz-Datenbank entscheidet eine Person, ob die Assemblierung vollständig ist oder ob zusätzliche Fragmente generiert und assembliert werden müssen.

werden. Ist die Person noch immer nicht mit den Ergebnissen des Experiments zufrieden, so werden die Bilder überprüft, die bei der Generierung der Fragmente erzeugt wurden und in Bild-Datenbanken (*Image DB*) gespeichert sind.[3] Diese Schleife wird solange durchlaufen, bis die Aktivität *human validation* einen entsprechenden Ausgabewert erzeugt, d.h., bis die Person mit dem Ergebnis der Assemblierung zufrieden ist. In diesem Fall erfolgt die Analyse der Sequenz, die aus Vergleichen mit anderen, bereits bekannten Sequenzen, sowie etwa der Suche nach biologisch wichtigen Eigenschaften besteht (z.B., ob Gene in der Sequenz enthalten sind).

3 Benutzung von FlowMark

Die Funktionalität von FlowMark teilt sich in zwei Module auf: *Buildtime* und *Runtime*. Die grundlegende Funktion des Buildtime Moduls ist die Modellierung von Prozessen, die des Runtime Moduls ist die kontrollierte Ausführung von Workflows, die zuvor (während der Buildtime) spezifiziert wurden.

3.1 Allgemein

Bei der Benutzung von FlowMark wurden die folgenden Funktionalitäten identifiziert und genauer untersucht:

- *Simulation:* Das Simulationswerkzeug spielt ein sehr wichtige Rolle im Modellierungsprozeß. Durch dieses Werkzeug kann man den modellierten Prozeß noch in der Modellierungsphase auf Korrektheit und Vollständigkeit hin überprüfen. Damit es möglich,

[3]Man beachte, daß bei dem Toplevel-Workflow die Datenspeicher aus Gründen der Übersichtlichkeit nicht dargestellt wurden. Details sind in [6] nachzulesen.

logische Fehler bereits vor der Laufzeit des Workflows zu erkennen und ggf. zu beheben.

- *Graphische Modellierung:* Die Hauptfunktionen des graphischen Modellierungswerkzeugs sind die Definition von Aktivitäten, Daten- und Kontrollfluß, die Spezifikation von Übergangsbedingungen zwischen Aktivitäten sowie die Anbindung externer Programme.

  Leider erlaubt dieses Werkzeug nicht die Spezifikation von Schleifen, was wir als eine starke Einschränkung betrachten — Schleifen müssen durch den Einsatz künstlicher *Blöcke* simuliert werden. In realen Anwendungen (insbesondere in den Naturwissenschaften) gibt es jedoch auch Workflows, die Schleifen besitzen (siehe Abb. 1). Insofern ist diese Eigenschaft von FlowMark eine echte Einschränkung, die für wissenschaftliche Anwendungen relevant ist; darüber hinaus sind auch in kommerziellen Anwendungen Workflows denkbar, die Schleifen enthalten.

- *Monitor:* Der Monitor stellt eine interessante Hilfe dar, um die Ausführung eines Workflows zu verfolgen und zu überwachen. Dieses Werkzeug ist leicht zu bedienen und kann zur Unterstützung der Kontrolle ablaufender Workflows eingesetzt werden. Diese Funktionalität ist sowohl in wissenschaftlichen als auch in kommerziellen Anwendungen interessant.

- *Graphische Benutzeroberfläche (GUI):* Der Umgang mit (den vielen notwendigen) Fenstern der Benutzeroberfläche gestaltet sich problematisch – oft muß auf die Fensterliste zurückgegriffen werden, um auf einzelne Fenster zugreifen zu können. Wir fanden einen Fehler in der Benutzerführung, der im folgenden beschrieben wird.

  - Es fehlt eine Standardisierung der Benutzerführung, wie beispielsweise zwischen den Fenstern der Funktion "Data Structures" und den Fenstern der Funktion "Programs". So müssen Veränderungen in der Spezifikation von Datenstrukturen explizit gesichert werden, während dies bei Veränderungen im Bereich der Programme nicht notwendig ist.
  - Wir fanden einen Fehler in der Funktion "Data Structures": Für jede Veränderung in einer Datastruktur muß man sichern, es gibt aber überhaupt keine Funktion "Sichern". Werden in dieser Funktion Veränderungen vorgenommen, so wird eine Meldung "Veränderung sichern?" erzeugt, so daß ein Sichern erst zu diesem Zeitpunkt möglich ist.

- *Audit Trail:* Obwohl das Konzept der *Audit Trails* eine sehr wichtige Funktionalität von FlowMark ist, die Informationen über ausgeführte Workflows bereitstellt, ist die Verwendung dieser Informationen nicht einfach. Um diese Informationen nutzbar zu machen, wäre es sinnvoll, den erzeugten Audit Trail (etwa) in eine relationale Datenbank einzuladen, um die notwendigen Informationen über die Ausführung des Workflows anschließend durch z.B. SQL Befehle abfragen zu können. Diese Funktionalität ist in FlowMark nicht vorhanden. Eine entsprechende Verwaltung dieser Informationen wäre

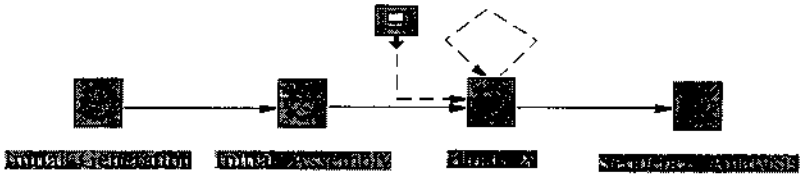

Abbildung 3: Toplevel-Workflow

aus unserer Sicht eine wünschenswerte Erweiterung von FlowMark, um die Ausführung von Workflows besser dokumentieren zu können.

- *Hilfefunktionen:* Nach unserer Meinung sind die Hilfefunktionen von FlowMark noch nicht befriedigend benutzerfreundlich. Bezüglich des Manuals und der Online-Hilfe sind die folgenden Punkte zu nennen:
  - Es sind leider nur wenige Beispiele vorhanden. Weitere Beispiele könnten die Darstellung veranschaulichen.
  - An einigen Stellen unklare bzw. fehlende Dokumentation. So wird nicht erklärt, wie ein DOS-Programm eingebunden werden kann – das Programm muß sich in einem bestimmten OS/2-Directory (OS2\MDOS) befinden; im Manual sind jedoch keine Informationen darüber enthalten.
- *Interoperabilität:* Es stellt sich heraus, daß FlowMark einen hohen Grad an Interoperabilität unterstützt. Durch sie kann man verschiedene Programme (C, C++, REXX) in verschiedenen Umgebungen (DOS, OS/2) ablaufen lassen. Das bedeutet, daß Programme oder sogar Altsysteme (*Legacy Systems*) in FlowMark eingebunden werden können.

3.2 Spezifikation wissenschaftlicher Workflows unter FlowMark

Nun beschreiben wir unsere Implementation des oben beschriebenen wissenschaftlichen Workflows "DNA-Sequencing" und gehen dabei insbesondere auf die Schwierigkeiten ein, die bei der Modellierung auftraten und z.T. zu einer Veränderung der originalen Repräsentation des Workflows (s. Abb. 1) führten.

Die Implementation des Toplevel-Workflows in FlowMark ist in Abbildung 3 dargestellt. Man beachte, daß die im originalen Workflow (s. Abb. 1) enthaltene Schleife in der FlowMark-Implementation nicht vorhanden ist, bzw., diese durch eine künstliche Aktivität BLOCK_X simuliert werden muß. Da FlowMark lediglich Schlingen (*self loops*) kennt, muß man die im orginalen Workflow vorhandene Schleife durch einen Block simulieren, d.h. die Aktivitäten der Schleife in einen Block zusammenfassen und diesen mit einer Schlinge versehen; BLOCK_X ist in Abbildung 4 dargestellt.

Man beachte, daß in Abbildung 1 *human validation* jeweils die erste und die letzte Aktivität der Schleife darstellt, d.h., es gibt nur einen Punkt, durch den man die Schleife betreten

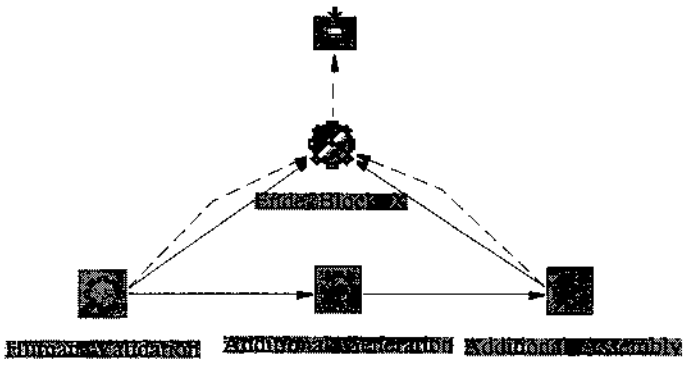

Abbildung 4: Body of Self-Loop (Block-X).

und verlassen kann. Diese Eigenschaft kann in FlowMark nicht auf natürliche Weise modelliert werden, da innerhalb von BLOCK_X eine lineare Liste von Aktivitäten erscheint, so daß sich in der graphischen Repräsentation des Workflows in FlowMark die Informationen über den Ablauf nicht widerspiegeln. Die Informationen über die Struktur des Workflows sind demgegenüber in den Übergangsbedingungen zwischen den Aktivitäten *human validation*, *additional assembly* und der künstlichen Aktivität *Ende_Block_X* "versteckt".

4 Fazit

FlowMark ist zur kontrollierten Ausführung von Workflows in Unternehmungen entwickelt und entsprechend ausgerichtet worden. In diesem Bericht haben wir untersucht, inwieweit dieses Produkt geeignet ist, Workflows in naturwissenschaftlichen Anwendungsgebieten zu unterstützen.

Diese Untersuchung wurde anhand eines stark strukturierten Workflows angestellt und nicht etwa bei Experimenten, bei denen erst während der Laufzeit über den weiteren Verlauf entschieden wird. Solche Workflows werden in jüngerer Zeit auch im Zusammenhang mit Business-Workflows unter dem Begriff *ad-hoc workflow* diskutiert. Ad-hoc Workflows sind insbesondere in wissenschaftlichen Anwendungen anzutreffen – allerdings ist die Unterstützung von Seiten kommerziell verfügbarer Workflow-Management Systeme in diesem Bereich recht schwach, und FlowMark stellt dabei keine Ausnahme dar. Die wesentlichen Erkenntnisse unserer Analyse können wie folgt zusammengefaßt werden:

- Modellierung von Schleifen ist nur unter Verwendung zusätzlicher, in der Anwendung nicht vorgesehener und daher als künstlich zu bezeichnender Konstrukte möglich. Die Beziehungen zwischen den Aktivitäten einer Schleife sind nicht graphisch darstellbar, sondern sind in den Übergangsbedingungen zwischen diesen Aktivitäten versteckt.
- Die Spezifikation und Verwendung von ad-hoc Workflows und die Wiederverwendung partieller Workflows gestaltet sich schwierig und wird vom System nicht unterstützt.

- Automatische Dokumentation ist eine für wissenschaftliche Anwendungen notwendige Funktionalität eines Workflow-Management Systems. Sie kann auf einem *Trace* der Aktivitäten basieren, die während einer Ausführung eines Workflows eintreten. Die Informationen werden von FlowMark in Form des *Audit Trail* gespeichert. Ein leichterer Zugang zu diesen Informationen wäre allerdings wünschenswert.

- Obwohl FlowMark ein Datenbanksystem zur Verwaltung workflow-relevanter Daten verwendet, kann man die Eigenschaften dieses Werkzeugs nicht benutzen, z.B., um komplexe Datenstrukturen (wie etwa Multimediadaten, Mengen oder Bäume) zu definieren.

FlowMark stellt ein für die konkreten Anforderungen von Geschäftsprozessen ausgerichtetes und diesbezüglich als weit entwickelt anzusehendes Produkt dar, das derzeit in vielen Unternehmen eingesetzt wird und dort zur Effizienzsteigerung der Geschäftsprozesse beiträgt. Bei unserer Analyse von FlowMark in einem neuartigen Einsatzbereich haben wir einige Beschränkungen identifiziert, die nicht nur für wissenschaftliche Workflows, sondern auch in anderen Einsatzbereichen von Workflow-Management relevant sind. Diesbezügliche Erweiterungen von Business-Workflow-Management Systemen können zu natürlicheren Modellierungen und schließlich einem vermehrten Einsatz dieser Systeme führen.

Literatur

[1] C. DeLisi. *The Human Genome Project.* American Scientist, 76:488–493, 1988.

[2] K.A. Frenkel. *The Human Genome Project and Informatics.* Communications of the ACM, 34(11):41–51, November 1991.

[3] IBM. *IBM FlowMark: Modeling Workflow, Rel 1.1.* Publ. No SH-19-8175-01, Sept. 1994.

[4] C.B. Medeiros, G. Vossen, and M. Weske. *WASA: A Workflow-Based Architecture to Support Scientific Database Applications.* In N. Revell and A.M. Tjoa (editors) Proceedings of the 6th DEXA Conference, Volume 978 of Springer *Lecture Notes in Computer Science*, pages 574–583, London, UK, September 1995.

[5] J. Meidanis. *FAT — A Fragment Assembly Toolkit.* In M. Vingron (editor) DIMACS Implementation Challenge - Fragment Assembly. 1995.

[6] J. Meidanis, G. Vossen, and M. Weske. *Using Workflow Management in DNA Sequencing.* Fachbericht Angewandte Mathematik und Informatik 23/95-I, Universität Münster, 1995.

[7] Strachan. *The Human Genome.* BIOS Scientific Publishers, 1992.

[8] M. Weske. *Survey of the Relationship between the Human Genome Project and Computer Science.* Fachbericht Angewandte Mathematik und Informatik 13/94-I, Universität Münster, 1994.