Lehrstuhl für Algorithm Engineering und Systemanalyse (LS XI)
Fachbereich Informatik
Universität Dortmund

Projektgruppenantrag für das Sommersemester 2006

21. November 2005

1 PG-Thema

Methoden der Computational Intelligence in der Bioinformatik

2 PG-Zeitraum

SoSe 06 und WS 06/07

3 PG-Umfang

8 SWS im ersten und zweiten PG-Semester, insgesamt 16 SWS

4 PG-Veranstalter

Dr. Thomas Bartz-Beielstein,
LS XI, JvF 20, Raum 2.67, Tel.: 9700-977
thomas.bartz-beielstein@udo.edu

Boris Naujoks,
LS XI, JvF 20, Raum 2.73, Tel.: 9700-369
boris.naujoks@udo.edu

Nicola Beume,
LS XI, JvF 20, Raum 2.73, Tel.: 9700-369
nicola.beume@udo.edu

5 PG-Aufgabe

5.1 Aufgabenbeschreibung

Bioinformatik ist ein interdisziplinäres Forschungsgebiet, das Biologie, Informatik, Mathematik und Statistik umfasst. Ziel ist die Erforschung und Entdeckung neuer biologischer Resultate wie z.B. die Entwicklung neuer Algorithmen und Modelle zur Verwaltung und zum Verständnis von riesigen Datenmengen sowie die Analyse und Interpretation von Aminosäuresequenzen oder Proteinstrukturen. In der modernen Medizin spielen Proteine eine wichtige Rolle, da mit ihrer Hilfe die Ursachen von Krankheiten erforscht, neue Medikamente entwickelt und deren Wirkungsweise entschlüsselt werden.

In der Projektgruppe sollen Lösungsverfahren mit Methoden der Computational Intelligence (CI) für typische Problemstellungen der Bioinformatik erarbeitet werden. Als ein Schwerpunkt ist die Optimierung der Topologie eines Hidden-Markov-Modells (HMM) mittels Evolutionärer Algorithmen vorgesehen.

Evolutionäre Algorithmen (EA), die sich an den Prinzipien der Evolution und der Genetik orientieren, haben sich in vielen praktischen Anwendungen als effiziente und robuste Optimierungsverfahren erwiesen. Ihre Bedeutung in der Bioinformatik wächst beständig [PBR05]. Die Universität Dortmund besitzt bei der Entwicklung von CI-Methoden eine weltweit herausragende Stellung, die durch die Einrichtung des Sonderforschungsbereichs 531 *Design und Management komplexer technischer Prozesse und Systeme mit Methoden der Computational Intelligence* belegt wird.

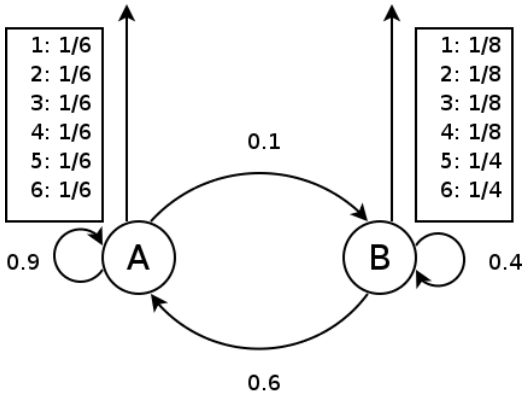

Abbildung 1: HMM mit 2 Zuständen. Zustand A beschreibt einen fairen Würfel, die Ausgabewahrscheinlichkeiten betragen für jede Zahl $1/6$. Mit einer Wahrscheinlichkeit von 10% wird ein unfairer Würfel (Zustand B) mit unterschiedlichen Ausgabewahrscheinlichkeiten benutzt. Die Wahrscheinlichkeit, vom Zustand B in den Zustand A zu wechseln, beträgt 60% .
In dieser PG sollen EA entwickelt werden, um die Struktur von HMM zu verbessern.

Hidden-Markov-Modelle (HMM) werden in der Bioinformatik eingesetzt, um DNA-Sequenzen oder Proteinstrukturen zu bestimmen. Ein anderes Anwendungsgebiet ist die Mustererkennung, beispielsweise zur Identifizierung von Worten innerhalb eines Sprachsignals. Ein HMM ist definiert als eine Menge von Zuständen, Ausgabe- und Übergangswahrscheinlichkeiten. Informationen über das interne (versteckte) Modell können nur über Ausgaben gewonnen werden. Eine Analyse der ausgegebenen Würfelergebnisse im Beispiel (s. Abb. 1) würde entlarven, dass das System nicht immer einen fairen Würfel benutzt.

Die Studierenden lernen in dieser Projektgruppe die Bioinformatik selbständig als Anwendungs- und Forschungsgebiet der Informatik kennen und werden dabei Kenntnisse über CI-Methoden (insbesondere EA), Softwaretechnologie, Datenverarbeitung sowie statistische Testverfahren erlangen, vertiefen und einsetzen. Neue Verfahren können international publiziert und die entstandenen Software-Komponenten als Erweiterung einer bestehenden Bibliothek veröffentlicht werden. Die Implementierung der Optimierverfahren und des HMM bauen auf vorhandene Java-Bibliotheken der Bioinformatik auf. Die Ergebnisse sollen außerdem graphisch dargestellt und statistisch validiert werden.

5.2 Problembeschreibung

Drei aufeinander aufbauende Probleme aus der Bioinformatik sollen in dieser PG untersucht werden.

5.2.1 Alignment und Vergleich von DNA-, RNA- und Proteinsequenzen

Ein *Alignment* ist eine Anordnung von zwei oder mehreren Sequenzen, um deren Unterschiede und Gemeinsamkeiten aufzuzeigen. Ein Alignment ist optimal, wenn es die meisten Übereinstimmungen und die geringsten Unterschiede zeigt. Der paarweise Vergleich von Protein- oder Aminosäuresequenzen ist eine fundamentale Aufgabenstellung der Bioinformatik, auf der viele andere Verfahren aufbauen (vgl. [DEKM98]). Standardmäßig kommt hier die dynamischen Programmierung zum Einsatz.

5.2.2 Vorhersage der Sekundärstruktur

Gene steuern biochemische Reaktionen, indem sie die Produktion von Enzymen kontrollieren. Enzyme gehören zur umfassenden Klasse der Proteine. Zum Verständnis der Genfunktion sind Kenntnisse der Proteinstrukturen notwendig. Proteine können als aus Aminosäuren bestehende Makromoleküle beschrieben werden. In Proteinen sind Aminosäuren durch Peptidbindungen miteinander verbunden. Die lineare Anordnung der Aminosäuren wird als *Primärstruktur* bezeichnet. Die *Sekundärstruktur* entsteht durch unterschiedliche Anziehungskräfte der Seitengruppen der Aminosäuren und kann beispielsweise eine Helix-Form darstellen. Diese dreidimensionale Struktur eines Proteins, bestimmt seine Funktion und ist allein

durch die Primärstruktur festgelegt. Durch die Kontrolle über die Primärstruktur können Gene die Funktion der Proteine beeinflussen.

Die Vorhersage der Sekundärstruktur aus der Primärstruktur ist eine wichtige Aufgabenstellung, da sich aus dieser Verhalten und Wirkungsweisen von Proteinen ableiten lassen. Für diese Aufgabenstellung wurden bereits Methoden des maschinellen Lernens herangezogen, z.B. neuronale Netze oder HMM.

Zu gegebenen Beobachtungen können die Übergangswahrscheinlichkeiten eines gegebenen HMM mit Methoden der dynamischen Programmierung bestimmt werden. Allerdings ist die optimale Struktur des zugrunde liegende HMM in der Regel nicht bekannt. Zum Auffinden besserer Strukturen können evolutionäre Algorithmen verwendet werden. Ein erster Ansatz wird in [Tho02] vorgestellt, [WHPBK05] stellen aktuelle, vielversprechende Ansätze dar. Die Einstellungen von EA, die bisher für Untersuchungen verwendet wurden, beruhen nur auf wenigen, vorläufigen Experimenten [Tho02, WHPBK05]. Dieser Bereich bietet offensichtlich ein großes Potenzial, um die Leistungsfähigkeit der Algorithmen zu steigern. Daher sollen Methoden aus dem Bereich der „Computational Statistics“ und dem „Design und der Analyse von Computerexperimenten“ zur Auswertung herangezogen werden, um die Bedeutung der gefundenen Resultate zu bewerten [BB05].

5.2.3 Beurteilung und Auswahl weiterer Verfahren

Mindestens eine weitere Problemstellung aus dem Bereich Bioinformatik, die mit CI-Methoden behandelt werden kann, soll von den PG-Teilnehmer in Abstimmung mit den Betreuern ausgewählt werden. Als Einstieg eignet sich der Artikel [PBR05]. Optional kann auch eine Problemstellung in Kooperation mit der Firma Protagen AG ausgewählt werden, so dass ein praxisrelevantes Aufgabengebiet bearbeitet wird.

5.3 Anwendungen

5.3.1 Alignment

Das Tutorium *How do I generate a pair-wise aligment* aus dem `BioJava`-Paket bietet einen guten Einstieg in die Problematik. Die darin beschriebene Methodik sollte durch die PG nachvollzogen und durch EA erweitert werden. Das `BioJava`-Paket steht kostenlos zur Verfügung `(www.biojava.org)`.

5.3.2 HMM Topologie Optimierung

Als Probleminstanzen sollen die in [Tho02] beschriebenen Datensätze aus der Protein Data Bank (PDB) und CASP Datensätze zum Lernen bzw. zur Validierung benutzt werden. Dieser Ansatz sollte in zwei Richtungen erweitert werden.

Algorithmisch: Unabhängig von den Problemstellungen aus dem Bereich der Bioinformatik (Sequenzanalyse, Strukturvorhersage) soll ein Basispaket mit bio-inspirierten Heuristiken in Java modelliert, implementiert und getestet werden. Hierzu zählen genetische Algorithmen, Evolutionsstrategien und Particle Swarm Optimization.

Problemspezifisch: Für die bio-inspirierten Heuristiken sollen problemspezifische Operatoren entwickelt werden, die zur Topologieoptimierung für HMM eingesetzt werden. Die Operatoren sollten zum `BioJava`-Paket kompatibel sein.

Ein leistungsfähiges Batch-System zur Durchführung berechnungsintensiver Versuche steht am LS11 zur Verfügung. Zur Visualisierung verschiedener Molekülstrukturen gibt es diverse frei verfügbare Pakete, die von der PG eingesetzt werden können. Das Göttinger `Bioinformatik-KNOPPIX` stellt eine entsprechende Sammlung von Tools bereit `(www.bioinf.med.uni-goettingen.de)`.

5.4 Hauptziel der Projektgruppe

Ziel ist die Modellierung, Implementierung und das experimentelle Testen (inklusive Analyse) evolutionärer Algorithmen zur Optimierung der Topologie von Hidden-Markov-Modellen. Die PG setzt HMM primär für die Vorhersage der Sekundärstruktur von Proteinen ein und analysiert die Ergebnisse graphisch und statistisch. Die Kompatibilität zum `BioJava`-Paket ermöglicht Vergleiche zu Verfahren anderer international anerkannter Forschergruppen. Die von der PG entwickelten Programme können in Abstimmung mit den Teilnehmern als Erweiterung des `BioJava`-Pakets zur Verfügung gestellt werden. Außerdem werden weitere Probleme der Bioinformatik optimiert und die resultierenden Verfahren auf umfassendere Problemklassen verallgemeinert.

6 Teilnahmevoraussetzungen

6.1 Notwendige Voraussetzungen

Die Teilnehmer der PG müssen als notwendige Voraussetzung mindestens in einem der Gebiete Computational Intelligence, Bioinformatik, grundlegende Statistikkenntnisse oder Softwaretechnologie (Testmethoden) Kenntnisse besitzen. Ferner muss eine objektorientierte Programmiersprache (z. B. C++ oder Java) beherrscht werden.

6.2 Wünschenswerte Voraussetzungen

- Computational Intelligence
- Bioinformatik
- Softwaretechnologie (Testmethoden)
- Grundlegende Statistikkenntnisse

7 Minimalziel

Es sollen mindestens zwei bio-inspirierte Verfahren zur Optimierung von HMM implementiert und getestet werden. Durch die Anbindung an `BioJava` besteht eine Schnittstelle zu verschiedenen Datenbanken und Standardbenchmarks, von denen mindestens zwei bearbeitet und mittels statistischer Auswertungen analysiert werden sollen.

Literatur

[BB05] Thomas Bartz-Beielstein. *Experimental Research in Evolutionary Computation – The New Experimentalism*. Natural Computing Series. Springer, Berlin, 2005. (im Druck).

[DEKM98] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[PBR05] S. K. Pal, S. Bandyopadhyay, and S. S. Ray. Evolutionary computation in bioinformatics: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part-C*, 2005. www.isical.ac.in/˜shubhra_r/ECBio.pdf.

[Tho02] René Thomsen. Evolving the topology of hidden markov models using evolutionary algorithms. In *Proceedings of Parallel Problem Solving from Nature VII (PPSN-2002)*, pages 861–870. Springer Verlag, 2002.

[WHPBK05] K. J. Won, T. Hamelryck, A. Prügel-Bennett, and A. Krogh. Evolving hidden markov models for protein secondary structure prediction. In *Proceedings of IEEE Congress on Evolutionary Computation*, pages 33–40, 2005.