Die Logik des Lebens

Zur Schlüsselrolle von KI-Methoden in der Biologie der Zukunft

Holger H. Hoos

Wie kaum eine andere Wissenschaft hat die Biologie in den letzten 10 Jahren von Entwicklungen in der Informatik profitiert. Dabei spielen Methoden und Ansätze aus der Künstlichen Intelligenz bereits heute eine Schlüsselrolle bei der Erforschung biologischer Prozesse und bilden ein wichtige Grundlage für die Erhebung und Analyse einer stetig wachsenden Menge von Daten mithilfe effizienter und massiv-paralleler experimenteller Ansätze. Anwendungen von KI-Methoden reichen von der Genomsequenzierung über die Funktionsbestimmung von Proteinen und Ribonukleinsäuren bis hin zur Modellierung komplexer Genregulations- und Stoffwechselvorgänge; hierbei kommen insbesonder Verfahren aus den Bereichen der heuristischen Suche und Optimierung sowie des maschinellen Lernens, aber auch Methoden aus der Wissensrepräsentation und -integration zur Anwendung. Die von vielen Experten erwarteten revolutionären Entwicklungen in der biologischen Grundlagenforschung, in der Medizin und in der Biotechnologie wird in hohem Maße auf dem Einsatz leistungsfähiger algorithmischer Verfahren zur Analyse, Modellierung und Simulation biologischer Prozesse und Systeme beruhen. Die sich in diesem Zusammenhang ergebenden Perspektiven und Anstöße haben ihrerseits das Potential, eine richtungsweisende Rolle in der KI des 21. Jahrhunderts zu spielen.

1 Einleitung und Überblick

Die Biologie beschäftigt sich mit einer der grundlegendsten Fragen, die Menschen sich seit Anbeginn wissenschaftlichen Denkens stellen: Wie funktioniert das Leben? Bezogen auf uns selbst hat diese Frage wichtige Verbindungen zur Medizin, wo Krankheiten als Störungen in unserem Körper und seinen Prozessen gesehen werden, deren Verständnis zumeist die Grundlage zur erfolgreichen Behandlung darstellt. Bezogen auf andere Lebewesen liefert die Biologie zunehmend wichtige Erkenntnisse hinsichtlich der Rolle von Organismen und Ökosystemen zur Stabilisierung oder zum Wandel wichtiger Aspekte unserer Umwelt.

Während Informatikmethoden und -anwendungen und die daraus erwachsenen Technologien innerhalb der letzten 50 Jahre unsere Gesellschaft und unser Leben auf fundamentale Weise verändert haben, beginnen die bahnbrechenden Fortschritte in der modernen Biologie erst jetzt, allgemein sichtbare Konsequenzen zu zeitigen. Es wird jedoch vielerseits erwartet, daß neue Erkenntnisse aus der modernen Biologie zu revolutionären Methoden zur Diagnose und Behandlung von Krankheiten führen werden, sowie zu einer Fülle von innovativen Anwendungen in der Biotechnologie, von effizienten Methoden zur nachhaltigen Nahrungsmittel- und Treibstoffgewinnung bis zur gezielten Konstruktion von Mikroorganismen zur Rohstoffverarbeitung und Schadstoffbeseitigung. Damit ist es sehr wahrscheinlich, daß die Biologie eine, vielleicht sogar *die* herausragende Rolle unter den Wissenschaften des 21. Jahrhunderts einnehmen wird. Eine Schlüsselfunktion im Hinblick auf die Realisierung dieses Potentials liegt dabei in der Verwendung von Methoden und Werkzeugen aus der Informatik, und insbesondere aus der KI.

Hauptziel dieses Beitrages ist es, die wichtige Rolle von KI-Methoden bei der Lösung biologischer Probleme zu skizzieren. Zu diesem Zwecke werden wir im folgenden eine kurze Einführung in einige der prominentesten Probleme in der modernen Molekularbiologie geben und die Rolle von KI-Methoden zur Lösung dieser zum größten Teil sehr schwierigen Probleme knapp umreißen. Aufgrund der regen Forschungsaktivitäten und der enormen Literatur in diesem Bereich wäre es aussichtslos, einen im Bezug auf die Probleme oder die entsprechenden Algorithmen vollständigen Überblick geben zu wollen. Aus diesem Grunde sind die ausgewählten Beispiele in erster Linie als Illustration der vielfältigen Verwendung einiger grundlegender KI-Methoden im Bereich der Bioinformatik zu verstehen, wobei die Auswahl in nicht unerheblichem Maße die Forschungsinteressen des Autors widerspiegeln. Im Anschluß an diesen Überblick wird dann die Rolle von KI-Methoden, und damit auch der KI-Forschung, in der Biologie des 21. Jahrhunderts erörtert.

2 Genome und Gene

Das Genom beinhaltet die gesamte genetische Information eines Organismus in Form von kettenförmigen Nukleinsäuresträngen, genauer gesagt, DNS. Die Sequenzierung von Genomen verschiedenster Organismen, von einfachen Bakterien bis zu Primaten, wie etwa dem Menschen, ist eine wichtige Komponente molekularbiologischer Forschung, und die daraus gewonnenen Informationen spielen eine zunehmend bedeutende Rolle für weite Bereiche der Biologie. Neben den Genen – Genomabschnitten, welche die Sequenzen funktionell wichtiger Proteine und Ribonukleinsäuren kodieren – umfassen Genome wichtige regulatorische und strukturelle Elemente, sowie in vielen Fällen eine große Anzahl von Sequenzen, deren Funktion noch immer unklar ist. Sowohl bei der Genomsequenzierung als auch im Zusammenhang mit der Analyse und Interpretation der so erhaltenen Information stellen sich eine Reihe von schwierigen Problemen, bei deren Lösung KI-Methoden eine wichtige Rolle spielen.

Die DNS-Stränge, welche die Genome eines Organismus bilden, sind in aller Regel zu lang für eine direkte Bestimmung der entsprechenden Sequenzen. Daher werden bei der Genomsequenzierung mehrfache Kopien eines langen DNS-Stranges in kleinere Fragmente zerlegt. Bei der sogenannten Schrotflintenmethode (engl. *"shotgun method"*) findet diese Zerlegung auf

zufällige Weise statt. (Dies ist die Methode, die bei der Sequenzierung des menschlichen Genoms von privatwirtschaftlicher Seite angewandt wurde.) Die hierbei enstehenden kurzen Sequenzabschnitte werden dann mit biochemischen Methoden sequenziert. Bereits dabei kommen relativ komplexe Datenanalyseverfahren zum Einsatz, um Ungenauigkeiten beim Auslesen der DNS-Sequenzen zu kompensieren. Die folgende Inferenz der Gesamtsequenz aus den einzelnen Fragmenten wäre allerdings ohne hocheffiziente Algorithmen auf der Ebene gesamter Genome völlig undenkbar. Diese Algorithmen basieren auf heuristischen Methoden zur Lösung einer Reihe von schwierigen Such- und Optimierungsproblemen, die sich in diesem Zusammenhang ergeben. Die Schwierigkeit dieser Probleme wird dabei in erheblichem Maße durch Sequenzierungsfehler und durch Wiederholungen innerhalb der Genomsequenz verursacht [3].

Eine wichtige Aufgabe in der Interpretation von Genomsequenzen besteht in der Bestimmung von Genen. Die Prozesse anhand derer die biomolekulare Maschinerie der Zelle Gene erkennt und die entsprechenden Produkte synthetisiert sind außerordentlich komplex und noch immer nicht vollständig verstanden. Bekannt ist jedoch, daß in diesem Zusammenhang eine Vielzahl verschiedener Motive und Sequenzmuster eine Rolle spielen; gleichzeitig besteht eine gewisse Robustheit gegenüber Veränderungen der Gene und der umgebenden genomischen Sequenzen.

Da der direkte Nachweis von Genprodukten in der Regel einen erheblich höheren experimentellen Aufwand erfordert als die Genomsequenzierung, bilden algorithmische Verfahren zur Identifikation von Genen eine wichtige Komponente in der Analyse und Annotation von Genomsequenzen. Die erfolgreichsten Algorithmen zur Lösung dieses schwierigen Genvorhersageproblems basieren auf der Integration von Verfahren zur Erkennung relativ einfacher Sequenzmerkmale mithilfe probabilistischer Modelle. Inbesondere kommen dabei sogenannte Verborgene Markow-Modelle (engl. *"hidden Markov models"*, kurz *HMMs*) zur Anwendung (siehe z.B. [7, 21]), welche auch eine wichtige Rolle in der Spracherkennung spielen.

Besondere Schwierigkeiten ergeben sich bei der Vorhersage von Genen höherer Organismen, in denen die informationstragende DNS-Sequenz von Abschnitten unterbrochen ist, die vor der Synthese des endgültigen Genprodukts entfernt werden. Diese sogenannten Introns finden sich beispielsweise in der überwiegenden Mehrzahl aller menschlichen Gene. Die Erkennung von Introns bereitet nach wie vor erhebliche Probleme bei der algorithmischen Genvorhersage, da wichtige Details des in der Natur für die Entfernung von Introns verantwortlichen Spleißmechanismus noch immer unbekannt sind. Ingesamt geht man davon aus, daß hierbei neben der Sequenz des zu spleißenden RNS-Transkripts auch dessen geometrische Struktur, und insbesondere die sogenannte Sekundärstruktur, welche Wechselwirkungen zwischen den Basen eines RNS-Moleküls beschreibt, eine wichtige Rolle spielt (siehe z.B. [13]). Damit ergeben sich bei der Intronvorhersage neben der Integration der verschiedenen Informationen auch molekulare Strukturvorhersageprobleme, auf die wir an späterer Stelle noch genauer eingehen.

Eine zunehmend bedeutendere Rolle bei der Genomanalyse spielen Verfahren, die Sequenzähnlichkeiten zwischen evolutionär verwandten Organismen ausnutzen. Solche Verfahren werden beispielsweise dazu verwendet, Prognosen zur Funktion neu gefundener (oder vorhergesagter) Gene zu gewinnen. Sie bilden auch die Grundlage phylogenetischer Inferenzmethoden, die dazu benutzt werden, evolutionäre Beziehungen zwischen Organismen zu bestimmen. Im Zusammenhang mit diesen letzteren Methoden ergeben sich zwei schwierige kombinatorische Probleme: Die Alinierung mehrerer DNS-Sequenzen (engl. *"multiple sequence alignment"*) sowie die Bestimmung von optimalen phylogenetischen Bäumen. Beide Probleme sind NP-schwer, aber von großer praktischer Bedeutung, und werden in der Praxis mithilfe von heuristischen Algorithmen gelöst. In beiden Fällen kommen dabei unter anderem lokale Suchverfahren zur Anwendung, welche auch in der KI eine wichtige Rolle spielen [12, 20]. Einige der erfolgreichsten Ansätze für die Konstruktion phylogenetischer Bäume basieren auf einer interessanten Kombination aus probabilistischen Modellen zur Sequenzevolution und heuristischen Suchmethoden zur Konstruktion wahrscheinlichkeitsmaximierender phylogenetischer Bäume (siehe z.B. [22]).

Ein weiteres wichtiges Problem im Bereich der Interpretation von Genomdaten ist die Identifikation von Sequenzmotiven, die beispielsweise entscheidende Rollen bei der Steuerung von Genaktivität spielen. Unbekannte Motive manifestieren sich in Form von relativ kurzen, sehr ähnlichen (aber in der Regel nicht völlig identischen) Sequenzabschnitten, welche sich z.B. in der Nähe von Genen finden. Das Aufspüren solcher Motive wird durch das zufällige Auftreten von Sequenzähnlichkeiten erschwert. Motividentifikationsprobleme können auf verschiedene Weise formalisiert werden, sind in der Regel aber NP-schwer und nur mithilfe heuristischer Verfahren lösbar. Dabei kommen neben speziellen Techniken zur robusten Behandlung statistisch schwach signifikanter Motive auch allgemeine Optimierungsverfahren, wie beispielsweise der im maschinellen Lernen oft verwendete EM-Algorithmus, zum Einsatz [6].

3 Biomolekulare Strukturen

Die Funktion von Biomolekülen steht in vielen Fällen in enger Verbindung zu deren geometrischer Struktur. Dies ist inbesondere der Fall für Proteine, welche als Biokatalysatoren Schlüsselrollen in unzähligen Reaktionen innerhalb der Zelle spielen. Proteine sind Kettenmoleküle, deren chemische Struktur durch eine Folge von Aminosäuren beschrieben werden kann. Diese sogenannte Primärstruktur ist einerseits durch Proteinsequenzierungsverfahren bestimmbar, andererseits aber in vielen Fällen auch direkt und auf einfache Weise aus genomischen DNS-Sequenzen ableitbar. Um ihre Funktion ausüben zu können, müssen die meisten Proteine eine klar bestimmte räumliche Struktur annehmen, welche durch Faltung der entsprechenden Aminosäurekette erreicht wird. Dabei enstehen lokale Strukturmotive, sogenannte Sekundärstrukturen, welche dann durch eine Vielzahl teilweise sehr komplexer Wechselwirkungen die Tertiärstruktur, d.h. die endgültige geometrische Form bilden.

Die Tertiärstruktur von Proteinen ist von großem und allgemeinem Interesse in vielen Bereichen der Biologie, Biochemie, Biotechnologie und Pharmakologie. Leider ist die experimentelle Bestimmung von Proteintertiärstrukturen äußerst aufwendig; daher ist für viele Proteine, deren Sequenz z.B. aus genomischen Daten bekannt ist, die Tertiärstruktur, wie auch in den meisten Fällen die genaue Funktion, unbekannt. Aus diesem Grunde sind algorithmische Methoden zur Bestimmung von Tertiärstrukturen auf der Basis von Sequenzdaten von großem Interesse, und die

entsprechenden Proteinstrukturvorhersageprobleme zählen zweifellos zu den wichtigsten Problemen der Bioinformatik.

Bei der Ab-Initio-Vorhersage wird die Tertiärstruktur von Proteinen ausschließlich auf der Basis physikalischer Prinzipien bestimmt. Dabei werden die dem Faltungsprozeß von Proteinen zugrundeliegenden Kräfte (inbesondere hydrophobische Effekte, aber auch elektrostatische und van-der-Waals-Kräfte, sowie Disulfitbindungen und Wasserstoffbrückenbildung) in Form einer Energiefunktion E modelliert, welche jeder Proteinstruktur S eine Energie $E(S)$ zuordnet. Die gesuchte Tertiärstruktur S^* hat nach gängigen thermodynamischen Annahmen in der Regel minimale Energie und kann daher durch globale Optimierung der Funktion E bestimmt werden. Dabei ergeben sich zwei grundlegende Schwierigkeiten: Zum einen ist das Verständnis der bei der Proteinfaltung wirkenden Kräfte unvollständig, wodurch sich Probleme bei der Konstruktion geeigneter Energiefunktionen ergeben. Zum anderen ist das Finden energieminimierender Strukturen ein äußerst komplexes Optimierungproblem; selbst für stark vereinfachte Proteinstrukturmodelle und Energiefunktionen ist dieses Problem NP-schwer [38].

Neben Informatikern, Bioinformatikern und Biochemikern beschäftigen sich auch theoretische Chemiker und Physiker sowie Forscher aus dem Bereich des *Operations Research* mit der Entwicklung und Analyse von Energieminimierungsverfahren zur Proteinstrukturbestimmung. Dabei kommt eine breite Palette von heuristischen Suchmethoden zur Anwendung, welche unter anderem Markowketten-Monte-Carlo-Verfahren [28], evolutionäre Algorithmen [39] und Ameisenalgorithmen [34] umfaßt. Die meisten dieser Verfahren basieren auf allgemeinen kombinatorischen Optimierungsverfahren, welche der Klasse der stochastisch-lokalen Suchmethoden zuzurechnen sind [18]. Diese Methoden und spezielle Algorithmen sind inbesondere innerhalb der letzten 10 bis 15 Jahre in der KI intensiv erforscht und angewandt worden, oftmals im Zusammenhang mit klassischen KI-Problemen aus den Bereichen Deduktion, Diagnose, Planen und Zeitablaufsteuerung (engl. *scheduling)*.

Bei der Lösung von Energieminimierungsproblemen in der Proteinstrukturvorhersage sind die von theoretischen Physikern und Chemikern entwickelten Verfahren oft deutlich verschieden von den in der KI und anderen Bereichen der Informatik betrachteten, sodaß ein erhebliches Synergiepotential besteht. Erhebliche Beiträge aus der KI sind dabei nicht nur in Form von konkreten Algorithmen zu erwarten, sondern auch im Bereich der Methoden, die bei der zielgerichteten Entwicklung und Analyse dieser heuristischen und zumeist hochgradig randomisierten Verfahren zum Einsatz kommen. Innerhalb der Informatik sind diese Methoden Forschungsgegenstand des noch jungen Gebiets der empirischen Algorithmik, welches historisch aber auch inhaltlich eng mit der KI verbunden ist.

Andere Ansätze zur Proteinstrukturvorhersage basieren auf den stetig wachsenden Beständen von experimentell bestimmten Proteinstrukturen. Beim sogenannten "Threading" werden Bewertungsfunktionen, die größtenteils auf statistischen Daten über einer Sammlung bekannter Strukturen basieren, im Zusammenhang mit diesen Strukturen selbst eingesetzt (siehe z.B. [37]). Große Mengen von kleinen Fragmenten bekannter Proteinstrukturen bilden in Form von sogenannten Rotamerbibliotheken die Grundlage für einige der leistungsfähigsten Proteinstrukturvorhersagemethoden [31]. Es ist zu erwarten, daß bei der weiteren Entwicklung solcher als wissensbasiert bezeichneten Verfahren KI-Methoden aus den Bereichen Datenintegration und heuristische Suche bzw. Optimierung sich as äußerst nützlich erweisen werden.

Neben der Proteintertiärstrukturvorhersage gibt es noch eine Reihe verwandter Probleme, zu deren Lösung KI-Methoden, insbesondere aus dem Bereich maschinelles Lernen, erfolgreich eingesetzt werden. Hierzu zählen unter anderem die Sekundärstrukturvorhersage [32], Probleme aus dem Bereich Proteinkomplexbildung und -interaktion (engl. *"docking"*) [14], sowie die Vorhersage von Protein-DNS- und Protein-RNS-Interaktionen [17].

Strukturvorhersageprobleme ergeben sich auch für Ribonukleinsäuremoleküle (RNS), welche nicht nur in der Proteinbiosynthese, sondern auch in der Gen- und Stoffwechselregulation und vielfältigen anderen zellulären Prozessen eine wichtige Rolle spielen (siehe z.B. [35]). Wie Proteine sind RNS-Moleküle Biopolymere deren Funktion in vielen Fällen von ihrer geometrischen Funktion bestimmt wird. Ebenso wie im Fall von Proteinen unterscheidet man bei RNS zwischen Primär-, Sekundär- und Tertiärstruktur. Allerdings spielt hier die Sekundärstruktur eine weitaus wichtigere Rolle für die Tertiärstrukturbildung sowie für direkte Wechselwirkungen mit anderen RNS- und DNS-Molekülen, sodaß sich die Forschung im Bereich der RNS-Struktur überwiegend auf die Sekundärstrukturebene konzentriert.

Analog zur Ab-Initio-Proteinstrukturvorhersage spielen auch in der RNS-Sekundärstrukturvorhersage Energieminimierungsansätze eine wichtige Rolle. Allerdings kann hier der wichtige Spezialfall der sogenannten pseudoknotenfreien Strukturen mithilfe dynamischer Programmierungsverfahren in polynomieller Zeit (genauer gesagt, $O(n^3)$) optimal gelöst werden [25]. Viele biologisch relevante RNS-Moleküle enthalten jedoch Pseudoknoten, und können damit von diesen Verfahren prinzipiell nicht behandelt werden. Weiterhin ist bekannt, daß das allgemeine Problem der Ab-Initio-Vorhersage von RNS-Sekundärstrukturen mit Pseudoknoten NP-schwer ist [24]; daher kommen in der Praxis (neben auf Spezialfälle beschränkten dynamischen Programmierungsalgorithmen) heuristische Suchverfahren zur Lösung dieses Problems zum Einsatz [16, 30], welche – wie auch im Fall der Proteinstrukturvorhersage – in engem Zusammenhang mit in der KI eingesetzten und untersuchten Methoden stehen. Im Vergleich zum Proteinstrukturvorhersageproblem ist hier allerdings das Spektrum bislang eingesetzter heuristischer Suchmethoden noch sehr beschränkt, sodaß erheblicher Raum für weitere Algorithmenentwicklung bleibt.

Das zur RNS-Sekundärstrukturvorhersage derzeit überwiegend eingesetzte Energiemodell basiert auf einer großen Anzahl von Parametern, welche die Energiebeiträge bestimmter Strukturmotive beschreiben. Die meisten dieser thermodynamischen Parameter wurden anhand von gezielten Messungen an speziell konstruierten, einfachen RNS-Molekülen experimentell bestimmt. Einige Parameterwerte wurden jedoch mithilfe eines einfachen evolutionären Algorithmus im Hinblick auf bessere Strukturvorhersagen optimiert [27]. Eine interessante und äußerst aktuelle Forschungsrichtung beschäftigt sich mit der Nutzung der wachsenden Menge experimentell bestimmter Sekundärstrukturen von komplexeren, biologisch relevanten RNS-Molekülen zur Konstruktion besserer Energiemodelle. Diese wissensbasierten Ansätze erfordern den Einsatz von leistungsfähigen Optimierungsfahren und können zum Teil auf Konzepte aus dem maschinellen Lernen aufbauen (siehe z.B. [11]). Ähnliche Methoden finden auch in der Entwicklung wissensbasierter Energiemodelle zur Protein-

strukturvorhersage Verwendung [42].

Im Zusammenhang mit der Struktur von RNS-Molekülen existieren eine Reihe weiterer Probleme, zu deren Lösung KI-Methoden zum Teil erheblich beitragen können (oder dies bereits tun). Hierzu zählt neben der RNS-Tertiärstrukturvorhersage [26] auch die Vorhersage von Interaktionen einerseits zwischen RNS-Molekülen [2] und andererseits zwischen RNS und DNS oder Proteinen [5].

Neben der Strukturvorhersage gibt es ein weiteres grundlegendes Problem, das sich mit der Struktur von Biomolekülen befaßt: das sogenannte Strukturdesignproblem. Dabei geht es um die zielgerichtete Konstruktion von Protein-, RNS- oder auch DNS-Molekülen mit bestimmten Struktureigenschaften. Solche maßgeschneiderten Moleküle sind von großem Interesse im Zusammenhang mit der Entwicklung neuer Wirkstoffe und Medikamente, aber auch in den relativ jungen Gebieten der biomolekularen Diagnostik und der Nanotechnologie.

Zum einen stellt sich dabei die Aufgabe, einzelne Proteine oder RNS-Moleküle mit einer bestimmten Struktur zu finden, d.h. Sequenzen zu bestimmen, die eine bestimmte Tertiärstruktur (oder, im Fall von RNS, Sekundärstruktur) annehmen und damit wichtige Wechselwirkungen mit anderen Molekülen eingehen können [9, 1]. Zum anderen ergibt sich das Problem, Mengen von Molekülen so zusammenzustellen, daß deren Wechselwirkung mit einer Reihe von anderen Molekülen vorgegeben Spezifizitäts- und Sensitivitätskriterien genügt. Dieses letztere Problem stellt sich zum Beispiel bei der Konstruktion von DNS-Mikrochips, welche eine zunehmend wichtige Rolle in der biologischen Grundlagenforschung, aber auch in der molekularen Diagnostik spielen (siehe z.B. [23]).

Die Lösung biomolekularer Designprobleme erfordert in der Regel hinreichend gute Verfahren zur Strukturvorhersage und in vielen Fällen darüberhinaus hocheffektive heuristische Suchverfahren. Damit spielen auch in diesem Bereich Methoden aus der empirischen Algorithmik und aus der KI eine wichtige Rolle.

4 Die Zelle und ihre Molekularen Mechanismen

Einige der interessantesten Anwendungen für KI-Methoden finden sich im Bereich der zellulären Systembiologie, einem jungen und äußerst aktiven Zweig der modernen Biologie. Im weitesten Sinne geht es hierbei um die Modellierung und Erforschung biologischer Systeme innerhalb der Zelle. Zu den aktuellen Forschungsthemen in diesem Gebiet zählen unter anderem die Modellierung von komplexen Genregulations-, Stoffwechsel- und Steuerungsprozessen.

Eine Schlüsselrolle kommt dabei relativ neuen, massivparallelen molekularen Diagnostikmethoden, wie etwa DNS-Mikrochip- und sogenannten ChIP-chip-Experimenten, aber auch modernen Chromatographieverfahren zu, die es ermöglichen, Zustände von komplexen biologischen Systemen auf der molekularen Ebene mit großer Effizienz zu charakterisieren. Die zunehmend breite Anwendung und ständige Weiterentwicklung dieser experimentellen Methoden im Zusammenhang mit rapiden Fortschritten in der Automatisierung von Laborabläufen führt zu einer enormen Menge von Daten, welche die Basis für diverse systembiologische Forschungsansätze bilden.

Schon die Organisation dieser Daten und des entsprechenden Wissens, erst recht jedoch die Analyse und Integration in größere Zusammenhänge bietet erhebliche Herausforderungen, deren Bewältigung in erheblichem Maße den Einsatz von KI-Methoden erfordert. Die hierbei angewandten Verfahren reichen von heuristischen Suchverfahren zur Graphalinierung (engl. *"graph alignment"*) [4] über Clusteranalyseverfahren zur Inferenz von Genregulationsbeziehungen [8] bis zu bayesschen Modellen zur Rekonstruktion von Stoffwechselwegen [15].

Zur Veranschaulichung betrachten wir ein prominentes Problem aus dem Bereich Genregulationsanalyse. Hierbei ist eine Reihe von Resultaten von DNS-Microchipexperimenten gegeben, von denen jedes einem bestimmten Zustand eines biologischen Systems entspricht. (Diese können z.B. auf von verschiedenen Patienten stammenden Proben basieren.) In jedem dieser Experimente werden gleichzeitig die Expressionswerte (d.h. mRNS-Mengen) für eine große Anzahl von Genen bestimmt, sodaß insgesamt für jedes untersuchte Gen ein Vektor von Expressionswerten vorliegt. Ziel ist es sodann, auf der Basis dieser Daten Gruppen von Genen zu identifizieren, die in gemeinsame regulatorische Prozesse involviert sind. Dies geschieht unter Verwendung von Distanz- oder Ähnlichkeitsmaßen auf den Expressionsvektoren.

Hierbei kommen typischerweise Verfahren zur Clusteranalyse zum Einsatz; diese reichen von relativ einfachen und effizienten agglommerativen Verfahren über den prominenten *K-Means*-Algorithmus – ein iteratives Verfahren, das mit einer festen Anzahl von Clustern arbeitet – bis zu komplexeren statistischen Methoden, wie etwa dem vielen maschinellen Lernverfahren zugrundeliegende EM-Algorithmus, zum Einsatz. (Ein Überblick über diese und andere zur Genexpressionsanalyse eingesetzten Clusteranalyseverfahren findet sich in [10].) Ein weiterer interessanter Ansatz ist die Coclusteranalyse (engl. *"biclustering"*), bei der gleichzeitig Gruppen von Genen und Experimenten bestimmt werden (siehe z.B. [29]). (Technisch gesehen geschieht dies durch simultane Selektion von Spalten- und Zeilenmengen in der durch die Genexpressionsvektoren gebildeten Expressionsmatrix.) Die hierbei zu lösenden Optimierungsprobleme sind in der Regel NP-schwer und erfordern daher den Einsatz heuristischer Methoden.

Ein weiterer zentraler Problembereich in der Systembiologie ist die Modellierung von komplexen Wechselbeziehungen, z.B. in der Genregulation. Aus der Sicht der KI-Forschung ergeben sich hierbei interessante Aufgabenstellungen in der Wissensrepräsentation und im Bezug auf geeignete Inferenzmethoden. Reine aussagenlogische Modelle (sogenannte *"Boolean networks"*), welche in diesem Zusammenhang bereits seit ca. 30 Jahren eingesetzt werden, haben sich trotz ihrer Einfachheit als überraschend nützlich zur Modellierung von Genregulationsmechanismen erwiesen [19].

Die inhärenten Beschränkungen dieses Ansatzes auf zwei Zustände und deterministische funktionale Wechselwirkungen können auf verschiedene Weise überwunden werden. Sogenannte probabilistische boolesche Netzwerke erlauben die explizite Repräsentation von Unsicherheit bzw. stochastischen, dynamischen Änderungen bezüglich der regulatorischen Wechselbeziehungen zwischen Genen [33]. Ähnliche Vorteile lassen sich durch die Verwendung von Bayesnetzen erzielen, wobei sich allerdings sehr schwierige Probleme bei der automatischen Konstruktion solcher Modelle auf der Basis von Genexpressionsdaten ergeben,

da die entsprechenden Strukturinferenzprobleme eine hohe Berechnungskomplexität aufweisen.

Eine detailliertere Repräsentation von Genregulationsbeziehungen wird durch den Einsatz von Systemen von Differentialgleichungen sowie von sogenannten additiven linearen Modellen ermöglicht (siehe z.B. [10]). Mit zunehmender Ausdrucksstärke der Modelle werden jedoch die entsprechenden Struktur- und Parameterinferenzprobleme in der Regel so schwierig (sprich: NP-schwer), daß sie nur unter Verwendung heuristischer Methoden gelöst werden können. Je nach Netzwerkgröße kann es hierbei äußerst nützlich sein, den Inferenzprozeß auf abstrahierte Regulationsbeziehungen auf der Basis von Clusteranalyseresultaten zu beschränken [40, 10].

Bei der automatischen Konstruktion von Modellen komplexer zellulärer Systeme und Wechselwirkungen ist es in zunehmendem Maße erforderlich, Daten aus verschiedenen experimentellen Analyseverfahren zu berücksichtigen. Zur Lösung der hierbei anfallenden Datenintegrationsaufgaben bieten sich verschiedene, zum Teil in der KI intensiv erforschte und angewandte Methoden an. Beispiele hierfür finden sich unter anderem in der Inferenz von Stoffwechselnetzwerken, welche Abhängigkeiten zwischen Enzymen über gemeinsame Stoffwechselprodukte bzw. zwischen Stoffwechselprodukten über auf diesen wirkenden Enzymen repräsentieren. Im Zusammenhang mit diesem Inferenzproblem ist kürzlich die Nützlichkeit von gewichteten Summen von gaußschen RBF-Kernen zur Integration von Genexpressions-, Lokalisations-, phylogenetischen Profil- und chemischen Kompatibilitätsdaten demonstriert worden [41], wobei allerdings die Gewichtsfaktoren manuell optimiert wurden. Hierbei werden zur Lösung des eigentlichen Inferenzproblems Verfahren aus dem kernbasierten überwachten Lernen angewandt, welche Wissen über Teile des zu konstruierenden Netzwerks ausnutzen.

Das wohl ehrgeizigste Ziel in der zellulären Systembiologie ist die Simulation ganzer Zellen auf biomolekularer Ebene. Es ist zu erwarten, daß derartige Simulationen zu gewaltigen Durchbrüchen nicht nur in vielen Bereichen der biologischen Grundlagenforschung, sondern auch in der Biotechnologie und Pharmakologie führen werden, da auf diese Weise Systemzustände und -reaktionen in bis dahin unbekannter Auflösung qualitativ und quantitativ dargestellt werden könnten. Weiterhin ließen sich auf diese Weise Rahmenbedingungen untersuchen, die im Labor schwer oder überhaupt nicht realisierbar sind.

Erste Simulationen für sehr einfache Zellen sind auf der Basis von regelbasierten Systemen, welche wichtige biomolekularen Reaktionen in kleinen Zeitschritten nachbilden, erstellt worden (siehe z.B. [36]). Entwicklungen in diesem Bereich sind noch immer in relativ frühen Anfangsstadien und werden in hohem Maße von Fortschritten in der Modellierung von Teilsystemen abhängen. Es ist jedoch klar, daß sich im Zusammenhang mit der Integration solcher Teilmodelle sowie bei der effizienten Simulation der relevanten Aspekte eines Gesamtmodells konzeptionelle und algorithmische Herausforderungen stellen, welche den in vielen Teilbereichen der KI anzutreffenden sehr ähnlich sind.

5 Ein Blick in the Zukunft

Wie in den vorausgehenden Abschnitten illustriert wurde, finden derzeit in vielen Bereichen der Molekularbiologie rasante Entwicklungen statt, die im wesentlichen durch Fortschritte auf dem Gebiet der experimentellen Labormethoden einerseits, und andererseits durch zunehmende Verwendung von Informatikmethoden ermöglicht werden. KI-Methoden spielen bereits heute eine Schlüsselrolle bei der Lösung der in diesem Zusammenhang anfallenden schwierigen algorithmischen Problemen. Selbstverständlich ist es schwierig, vielleicht sogar unmöglich, wichtige Entwicklungen in einem derartig dynamischen Bereich vorauszusehen; dennoch gibt es absehbare Tendenzen, Ziele und Zukunftsvisionen, welche wir im folgenden kurz, und zugegebener Weise zum Teil recht spekulativ umreißen wollen.

Es ist klar abzusehen, daß weitgehend automatisierte, massivparallele experimentelle Analysemethoden in zunehmend Maße alle Bereiche biologischer Forschung durchdringen werden, und sich darüberhinaus auch in medizinischen und biotechnologischen Anwendungen in der nächsten Zukunft fest etablieren werden. Dies führt zur Erfassung und Speicherung großer Datenmengen, deren Analyse den Einsatz von KI-Methoden im breiten Rahmen erfordern wird. Neue Herausforderungen für die Bioinformatik und die KI werden sich dabei zum einen aufgrund der Einführung neuer Technologien, etwa im Bereich der Genomsequenzierung oder Proteinidentifikation, stellen, zum anderen bei der Integration von Wissen aus einem breiten Spektrum von Daten unterschiedlicher Art.

Gleichzeitig ist es leicht vorstellbar, daß der immer leichtere Zugang zu riesigen Mengen detaillierter Informationen über Komponenten und Zustände biologischer Systeme dazu führen kann, daß bereits in der nahen Zukunft der überwiegende Teil molekularbiologischer Forschungsprojekte in erster Linie statt im Labor am Rechner durchgeführt wird und die Datenanalyse in den Vordergrund biologischer Forschungsaktivität tritt. An diesem Punkt würden statt umständlicher und kostspieliger Laborexperimente rechnergestützte Analysemethoden zum limitierenden Faktor bezüglich des wissenschaftlichen Fortschritts in weiten Bereichen der Biologie werden, sodaß eine der wichtigsten Aufgaben in der Entwicklung (und Anwendung) neuer Algorithmen bestünde. Mit der zunehmenden Integration der Bioinformatik in die Biowissenschaften ist zu erwarten, daß zum einen große Veränderungen in der wissenschaftlichen Ausbildung von Forschern in diesem Bereich nötig werden, zum anderen jedoch vielgestaltige Perspektiven für Forscher aus der Informatik, und insbesondere aus der KI entstehen.

Im Bereich der Molekularbiologie zeichnen sich eine Reihe interessanter Entwicklungen ab. Zum einen besteht eine klare Tendenz zur Integration von Wissen zur Lösung nahezu aller grundlegenden Probleme, von der Genomanalyse bis hin zur Proteinstrukturvorhersage. Viele Forscher gehen davon aus, daß in einigen Bereichen in absehbarer Zukunft eine kritische Schwelle erreicht werden könnte, jenseits derer viele Probleme allein aufgrund der verfügbaren Datenmenge qualitativ sehr viel besser lösbar werden, als dies bislang der Fall ist. Dies gilt insbesondere für die Vorhersage von molekularen Strukturen und Interaktionen, aber auch für die Genvorhersage und bestimmte Motividentifikationsprobleme. Etwas unklar ist dabei, in welchem Maß sich das Skalierungsverhalten der Algorithmen, die auf diesen Daten operieren müssen, als limitierender Faktor herausstellen könnte.

Es ist davon auszugehen, daß systembiologische Ansätze bereits in der näheren Zukunft eine entscheidende Rolle innerhalb der biologischen Forschungsgemeinde spielen werden. Auch wenn das Ziel, eine gesamte Zelle im Detail zu simulieren, si-

cherlich noch in weiter Ferne liegt, ist realistisch zu erwarten, daß Teilsysteme von zunehmender Komplexität innerhalb der nächsten 5-10 Jahre hinreichend detailliert modelliert und simuliert werden können, daß ein großes Spektrum wesentlicher neuer Erkenntnisse über die Funktionsweise der entsprechenden Systeme gewonnen werden kann. Dabei ist es gut vorstellbar, daß KI-Methoden und -Ansätze, die in diesem Bereich bereits heute eine wichtige Rolle spielen, unverzichtbare Beiträge bei der Konstruktion und Nutzung dieser Modelle leisten werden.

Selbstverständlich beschränken sich die rapiden Fortschritte auf der Basis neuer experimenteller und algorithmischer Methoden nicht auf den Bereich der Molekularbiologie. In der Tat hat der zunehmende Einsatz von Methoden aus der Molekularbiologie und der Bioinformatik nach Ansicht vieler Wissenschaftler bereits eine Revolution weiter Bereiche der Mikrobiologie, Botanik, Zoologie, Ökologie und Humanbiologie ausgelöst. Dabei ist zu erwarten, daß sich in diesen Bereichen in zunehmendem Maße formale Modelle oberhalb der biomolekularen Ebene etablieren werden, welche beispielsweise die Funktion von Organen oder Ökosystemen auf einer höheren Abstraktionsebene beschreiben. Die Integration von Modellen und Modellkomponenten auf verschiedenen Abstraktionsebenen wird dabei faszinierende und schwierige Herausforderungen stellen, zu deren Bewältigung KI-Ansätze prinzipiell höchst geeignet erscheinen.

Von größtem allgemeinen Interesse sind in jedem Falle zukünftige Entwicklungen in den Anwendungsbereichen der modernen Biologie. In der Medizin werden derzeit Ansätze zur molekularen Diagnose und Therapeutik intensiv erforscht, und es ist zu erwarten, daß einige dieser Methoden bereits innerhalb der nächsten 5 Jahre zur Anwendungsreife gelangen. Es ist gut vorstellbar, daß innerhalb von 10-15 Jahren Systeme zur frühen und genauen Diagnose einer Vielzahl von Erkrankungen auf der Basis massiv-paralleler molekularbiologischer Verfahren, etwa zur Genexpressions- oder Proteomanalyse, routinemäßig eingesetzt werden. Ähnliche Daten werden mit großer Wahrscheinlichkeit in Kombination mit herkömmlichen diagnostischen Daten auch eine wichtige Rolle bei der detailierten Überwachung von Krankheits- und Behandlungsverläufen spielen.

Eine weitere mögliche Entwicklung mit weitreichend Konsequenzen liegt in der Nutzung individueller Genomdaten (im Extremfall der kompletten Genomsequenz) zu Zwecken der Abschätzung von Erkrankungsrisiken sowie zur Prognose von Krankheits- und Heilungsverläufen. Dies ist inbesondere im Bezug auf sogenannte multifaktorielle Erkrankungen relevant, an deren Entstehung sowohl genetische als auch umweltbezogene Faktoren beteiligt sind. In diesem Zusammenhang ergeben sich potentiell erhebliche ethische Probleme, da derartige Informationen über individuelle Risiken weitreichende Konsequenzen für die Lebensumstände der betroffenen Menschen haben können (z.B. in Form stark differenzierter Krankenversicherungsbeiträge).

Wichtige Fortschritte sind auch in der Methodik zur Entwicklung von Medikamenten zu erwarten. Hier ist davon auszugehen, daß Verbesserungen in der molekularen Strukturvorhersage im Zusammenspiel mit detaillierten Modellen wichtiger molekularer Mechanismen ein zunehmend zielgerichtetes Design von Wirkstoffen mit minimalen Nebenwirkungen ermöglichen wird. Eine andere äußerst interessante Perspektive besteht im Bezug auf kombinierte Wirkstoffsystemen, deren Aktivität von molekularen Diagnosen innerhalb individueller Zellen gesteuert wird. Basierend auf detaillierten Kenntnissen zellulärer Mechanismen und der Fähigkeit, in die zugrundeliegenden molekularen Wechselwirkungen gezielt einzugreifen, könnten solche Systeme inbesondere die Behandlung von Krebserkrankungen revolutionieren. Darüberhinaus ergeben sich interessante und wahrscheinlich sehr weitreichende Anwendungsmöglichkeiten in der molekularen Gentherapie, insbesondere im Hinblick auf die (ethisch allerdings kontroverse) Durchführung von genetischen Reparaturen an Keimzellen zur Vermeidung von Erbkrankheiten.

Eine der gewagtesten Visionen ist die des „Lebens vom Reißbrett“, die in ihrer extremsten Variante die komplette Neukonstruktion von Lebewesen zum Ziel hat. Obwohl genetisch manipulierte Organismen bereits in großem Maßstab hergestellt und genutzt werden (insbesondere in der Nahrungsmittelproduktion), ist davon auszugehen, daß der Schritt zur vollständigen Neukonstruktion zumindest für die nächsten 10 Jahre noch außerhalb unserer Reichweite liegen wird. Gleichzeitig ist es durchaus vorstellbar, und in gewisser Hinsicht sogar recht wahrscheinlich, daß die wissenschaftlichen Vorraussetzungen für diesen Schritt in den nächsten 25-50 Jahren gegeben sein könnten. Während es keiner großen Phantasie bedarf, sich zahlreiche nützliche Anwendungen solcher maßgeschneiderten Organismen vorzustellen, wird die Frage nach der Verantwortbarkeit und ethischen Vertretbarkeit solcher Entwicklungen wohl sehr schwierig zu beantworten sein.

In jedem Falle wird die Entwicklung und Anwendung geeigneter Informatikmethoden von entscheidender Bedeutung für die Realisierung dieser Perspektiven und Visionen sein. Aufgrund der Natur der hierbei zu lösenden algorithmischen Probleme kommt dabei Verfahren und Ansätzen aus der KI eine Schlüsselrolle zu, und für die KI-Forschung ergibt sich in diesem Zusammenhang ein breites Spektrum faszinierender Herausforderungen.

Literatur

[1] M. Andronescu, A.P. Fejes, F. Hutter F, H.H. Hoos, A. Condon. A new algorithm for RNA secondary structure design. Journal of Molecular Biology, 336(3): 607–624, 2004.

[2] M. Andronescu, Z.C. Zhang, A. Condon. Secondary Structure Prediction of Interacting RNA Molecules. Journal of Molecular Biology 345(5): 987–1001, 2005.

[3] S. Batzoglou. Algorithmic Challenges in Mammalian Genome Sequence Assembly. In: Encyclopedia of genomics, proteomics and bioinformatics. John Wiley and Sons, 2005.

[4] J. Berg, M. Lässig. Local graph alignment and motif search in biological networks. PNAS 101(41): 14689–14694, 2004.

[5] N. Bhardwaj, R.E. Langlois, G. Zhao, H. Lu. Kernel-based machine learning protocol for predicting DNA-binding proteins. Nucleic Acids Research 33(20): 6486–93, 2005.

[6] J. Buhler, M. Tompa. Finding motifs using random projections. Journal of Computational Biology, 9(2): 225–242, 2002.

[7] C. Burge, S. Karlin. Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology, 268, pp. 78–94, 1997.

[8] J.M. Claverie. Computational methods for the identification of differential and coordinated gene expression. Human Molecular Genetics, 8(10): 1821–1832, 1999.

[9] B.I. Dahiyat, S.L. Mayo. De novo protein design: fully automated sequence selection. Science 278(5335): 82–87, 1997.

[10] P. D'haeseleer, S. Liang, R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16(8): 707–726, 2000.

[11] Y. Ding. Statistical and Bayesian approaches to RNA secondary structure prediction. RNA 12(3): 323–331, 2006.

[12] R.C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5): 1792–1797, 2004.

[13] V. Goguel V, M. Rosbash. Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. Cell 72(6): 893–901, 1993.

[14] J.J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C.A. Rohl, D. Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of Molecular Biology 331(1): 281–299, 2003.

[15] M.L. Green, P.D. Karp. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinformatics 5:76, 2004.

[16] A.P. Gultyaev, F.H.D. van Batenburg, C.W.A. Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. Journal of Molecular Biology 250(1): 37–51, 1995.

[17] L.Y. Han, C.Z. Cai, S.L. Lo, M.C. Chung, Y.Z. Chen. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA 10(3): 355–368, 2004.

[18] H.H. Hoos, T. Stützle. Stochastic Local Search: Foundations and Applications. Morgan Kaufmann Publishers/Elsevier, 2004.

[19] S. Huang. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. Journal of Molecular Medicine 77(6): 469–480, 1999.

[20] J.M. Keith, P. Adams, M.A. Ragan, D. Bryant. Sampling phylogenetic tree space with the generalized Gibbs sampler. Molecular Phylogenetics and Evolution 34(3): 459–468, 2005.

[21] A. Krogh. Two methods for improving performance of an HMM and their application for gene finding. Proc. of 5th Int. Conf. on Intelligent Systems for Molecular Biology, pp. 179–186, AAAI Press, 1997.

[22] P.O. Lewis. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Molecular Biology and Evolution 15(3): 277–283, 1998.

[23] F. Li, G.D. Stormo. Selection of optimal DNA oligos for gene expression arrays. Bioinformatics 17(11): 1067–76, 2001.

[24] R.B. Lyngsø, C.N.S. Pedersen. RNA pseudoknot prediction in energy-based models. Journal of Computational Biology 7(3): 409–427, 2000.

[25] R.B. Lyngsø, M. Zuker, C.N.S. Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction. Bioinformatics 15(6): 440–445, 1999.

[26] F. Major, S. Lemieux, M. Ftouhi. Computer RNA Three-Dimensional Modeling from Low-Resolution Data and Multiple-Sequence Information. In: Molecular Modeling and Structural Determination of Nucleic Acids American Chemical Society Books, pp. 394–404, 1998.

[27] D.H. Mathews, J. Sabina, M. Zuker, D.H. Turner. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. Journal of Molecular Biology 288: 911–940, 1999.

[28] A. Mitsutake, Y. Sugita, Y. Okamoto. Generalized-Ensemble Algorithms for Molecular Simulations of Biopolymers. Biopolymers (Peptide Science) 60(2): 96–123, 2001.

[29] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22(9): 1122–1129, 2006.

[30] J. Ren, B. Rastegari, A. Condon, H.H. Hoos. HotKnots: Heuristic Prediction of RNA Secondary Structures Including Pseudoknots. RNA 11(10): 1494–1504, 2005.

[31] C.A. Rohl, C.E. Strauss, K.M. Misura, D. Baker. Protein structure prediction using Rosetta. Methods in Enzymology 383: 66–93, 2004.

[32] B. Rost, C. Sander. Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology 323(2): 584–599, 1993.

[33] I. Shmulevich, E.R. Dougherty, S. Kim, W. Zhang. Probabilistic Boolean networks: a rulebased uncertainty model for gene regulatory networks. Bioinformatics 18(1): 261–274, 2002.

[34] A. Shmygelska, H.H. Hoos. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. BMC Bioinformatics 6:30, 2005.

[35] R.W. Simons, M. Grunberg-Manago. RNA structure and function (editors). Cold Spring Harbor Laboratory Press, 1998.

[36] M. Tomita. Whole-cell simulation: a grand challenge of the 21st century. TRENDS in Biotechnology 19(6): 205–210, 2001.

[37] A.E. Torda. Protein Threading. In: The Proteomics Handbook, Humana Press, pp. 921–938, 2005.

[38] R. Unger, J. Moult. Finding the lowest Free-Energy Conformation of a protein is an NP-hard problem - Proof and Implications. Bulletin of Mathematical Biology, 55(6): 1183–1198, 1993.

[39] R. Unger, J. Moult. Genetic algorithms for protein folding simulations. Journal of Molecular Biology, 231(1): 75–81, 1993.

[40] M. Wahde and J. Hertz. Coarse-grained reverse engineering of genetic regulatory networks. Biosystems 55(1-3): 129–136, 2000.

[41] Y. Yamanishi, J.-P. Vert, M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics, Vol. 20, Suppl. 1, pp. i363–i370, 2004.

[42] J. Zhang, R. Chen, J. Liang. Empirical Potential Function for Simplified Protein Models: Combining Contact and Local Sequence-Structure Descriptors. PROTEINS: Structure, Function, and Bioinformatics 63: 949–960, 2006.

Kontakt

Prof. Holger H. Hoos
Department of Computer Science
University of British Columbia
2366 Main Mall
Vancouver, BC, V6T 1Z4, Canada

Tel.: +1 604 822-1964
Fax: +1 604 822-5485
E-Mail: hoos@cs.ubc.ca

Bild

Holger Hoos ist *Associate Professor* an der University of British Columbia in Vancouver (Kanada), wo er als Mitgründer des *Bioinformatics, Empirical and Theoretical Algorithmics Laboratory (BETA Lab)* vorwiegend Forschung im Bereich empirische Algorithmik (mit Anwendungsschwerpunkten in der Bioinformatik und in der KI) betreibt.