Universität zu Lübeck

Bioinformatik – Informationsverarbeitung in der Biologie

Prof. Dr. Thomas Martinetz

Institut für Neuro- und Bioinformatik

Sonderdruck aus FOCUS MUL 18, Heft 2 (2001)

Bioinformatik – Informationsverarbeitung in der Biologie

Th. Martinetz

Die Bioinformatik ist derzeit in der Forschungspolitik, aber insbesondere auch in den Medien, ein sehr aktuelles Thema. Es vergeht kaum eine Woche, in der nicht irgendwo eine Professur für Bioinformatik ausgeschrieben wird. Die entsprechenden Universitäten kämpfen um geeignete Kandidaten, und manche der Professuren werden unbesetzt bleiben, weil die Nachfrage zu groß ist.

Was ist der Grund für das enorme Interesse an der Bioinformatik? Warum werden mehrere 100.000.000 DM in den nächsten Jahren in diese Forschungsrichtung investiert? Hoffnungen auf einen hohen „Return of Investment“ wurden vor allem ausgelöst durch die erfolgreich klingenden Meldungen und den begleitenden Medienrummel zum „Human Genom Project“, die Sequenzierung des menschlichen Genoms. Der Schlüssel zur Heilung vieler Krankheiten liegt im Verständnis molekulargenetischer Prozesse. Dies betrifft das höchste Gut aller Menschen, und dafür ist man bereit, viel Geld auszugeben. Es ist jedoch auch eine erhöhte Skepsis gegenüber vielen dieser Heilsversprechungen geboten, und vieles, was als Vision verkauft wird, wird Utopie bleiben. Die derzeit aktuelle Ethikdiskussion zur Biotechnologie wird hier hoffentlich zur Klärung überzogener Erwartungen und Befürchtungen in der Bevölkerung beitragen (siehe Beitrag Dr. Maio). Die Integrität der Wissenschaftler wird angesichts der Möglichkeit, durch vielversprechende Zukunftsvisionen enorme Mengen an Fördergeldern in ihre Fachrichtungen zu lenken, auf eine harte Probe gestellt.

Unbestritten ist, dass der Fortschritt in der Biotechnologie wesentlich durch den Einsatz von Computern getragen wird. Die Entschlüsselung des menschlichen Genoms konnte nur durch den Einsatz von Höchstleistungsrechnern gelingen. Clevere Algorithmen waren erforderlich, die enorme Flut von Daten zu sortieren und mosaikartig zum Gesamtbild zusammenzusetzen. Die Firma Celera, die als privates Unternehmen eingestiegen war, um in Konkurrenz zum öffentlich geförderten weltweiten „Human Genome Project“ im Alleingang das menschliche Genom zu sequenzieren, konnte nur konkurrieren, weil es mit komplexeren Algorithmen an die Problemstellung heranging. Diese Algorithmen ermöglichten es auf clevere Art und Weise, mit bruchstückhafteren Daten auszukommen und in dieser Hinsicht den Verfahren, die im öffentlich geförderten „Human Genome Project“ angewandt wurden, überlegen zu sein. Es war also die Informatik, die hier ganz wesentlich den Ausgang des Rennens bestimmte.

Die Biologie, insbesondere die Molekularbiologie, braucht die Informatik, und dies noch mehr in der Zukunft. Unternehmen wie Celera produzieren molekularbiologische Daten in industriellem Maßstab. Hier sind ganze „Produktionsstraßen“ zur „Herstellung“ dieser Daten im Einsatz, und Computer und Software, wie man sie aus Produktionsbetrieben kennt, werden verwendet. Insbesondere werden aber Computer und entsprechende Software zur Speicherung und Nutzung dieser Daten benötigt. Die Menge dieser Daten hat sich in den vergangenen Jahren bereits alle drei Jahre vervierfacht. Dies wird sich noch steigern, insbesondere durch den zunehmenden Einsatz der Genchiptechnologie. Damit wird die Menge der Daten sogar schneller ansteigen als die Leistungsfähigkeit der Computer, womit eine problematische Kluft entsteht. Auf der anderen Seite häufen sich Meldungen, dass viele der produzierten Daten noch sehr fehlerhaft sind. Die Auswertemethoden müssen also noch erheblich verbessert werden, und Daten müssen im großen Maßstab auf Konsistenz geprüft werden. Auch dazu wird die Informatik dringend gebraucht.

Zukünftig wird die Informatik jedoch auch als Grundlagenwissenschaft informationsverarbeitender Prozesse an zentraler Stelle zum Verständnis molekularbiologischer Abläufe benötigt. Denn ein wirkliches Verständnis der nun vorliegenden Genomsequenzen verlangt das Verständnis der darauf basierenden komplizierten regulatorischen Netzwerke, der Wechselwirkungsnetzwerke zwischen DNA und Proteinen sowie Proteinen untereinander. Erst mit dem Verständnis dieser hochkomplexen Regulationsmechanismen wird das Potenzial, das die Sequenzierung des menschlichen Genoms verspricht, nutzbar gemacht werden können. Diese regulatorischen Netzwerke in ihrer Gesamtheit, mit dem Genom lediglich als ein Element dieser Netzwerke, bestimmen letztendlich die Expression phänotypischer Aspekte. Zum Verständnis dieser hochkomplexen Mechanismen braucht man die Informatik nicht nur als Lieferant von cleveren Algorithmen und schnellen Supercomputern. Man wird die Informatik vor allem auch als Grundlagenwissenschaft der Informationsverarbeitung benötigen, denn es sind letztendlich Informationsverarbeitungsprozesse, die solch hochkomplexe regulatorische Systeme determinieren.

Mit hochvernetzten Systemen beschäftigt sich ein anderer, komplementärer Zweig der Bioinformatik schon seit langem. Unter Bioinformatik versteht man derzeit vor allem „Informatik für die Biologie“, den Einsatz und die Entwicklung von Algorithmen und Computern zur Unterstützung des Verständnisses und der Nutzung molekulargenetischer Prozesse. Ich nenne dies Molekulare Bioinformatik. Die Bioinformatik in seiner gesamten Breite beinhaltet jedoch noch ein zweites Teilgebiet, die „Biologie für die Informatik“. Was kann die Informatik von der Informationsverarbeitung in der Biologie, in biologischen Systemen, lernen? Welche Prinzipien der Informationsverarbeitung werden zum Beispiel im Nervensystem genutzt? Oder auf welchen informationsverarbeitenden Prozessen basiert die biologische Evolution? Die Evolution generiert Information, Information über die Umwelt, die wiederum im Genom gespeichert wird. Informationsinhalte werden adaptiert durch Mutation und optimiert durch Selektion, und dies offensichtlich überaus erfolgreich. Biologische Systeme sind von enormer Komplexität. Um solche Systeme hervorzubringen, aufzubauen, zu organisieren und in einer komplexen Umwelt zu steuern, bedarf es besonderer Informationsverarbeitungsprinzipien. Viele Aufgaben der Informationsverarbeitung können von biologischen Systemen immer noch um Größenordnungen besser bewältigt werden als von den schnellsten heutigen Computern. Ein grundlegendes Verständnis dieser informationsverarbeitenden Prinzipien wäre sowohl von hoher Erkenntnis- als auch Praxisrelevanz. Immer wieder geht es dabei um hochvernetzte Systeme, am deutlichsten bei der Informationsverarbeitung im Nervensystem, mit dem sich die Neuroinformatik als Teilgebiet der Bioinformatik beschäftigt. Ein Neuron ist im Mittel mit 10.000 anderen Neuronen verschaltet. Interessanterweise sind die mathematischen Gleichungen, die die Dynamik solch hochvernetzter Systeme beschreiben, ob neuronale Netzwerke oder regulatorische Netzwerke, immer von ähnlicher Struktur. Auf dieser grundlegenden Ebene treffen sich die beiden Teilgebiete der Bioinformatik, die „Informatik für die Biologie“ und die „Biologie für die Informatik“, und können maßgeblich voneinander profitieren.

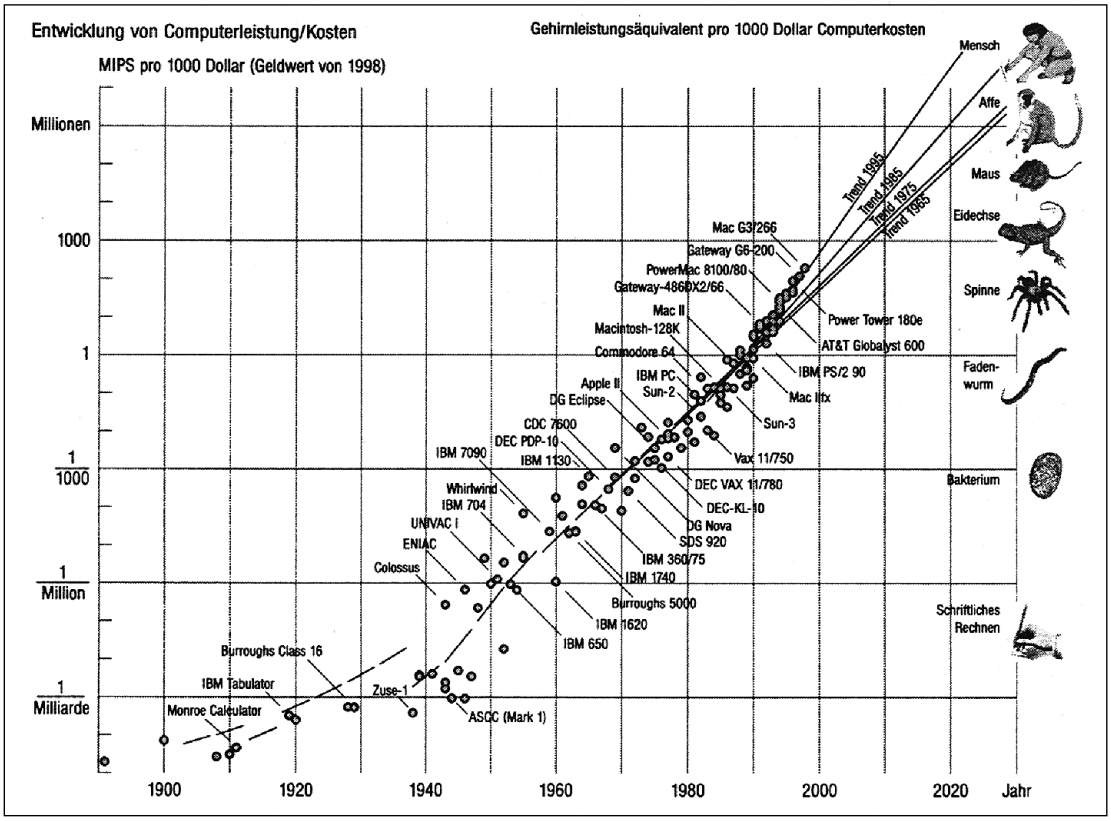

Abb. 1: Die Rechenleistung von Computern in Millionen Instruktionen pro Sekunde (MIPS) pro 1000 Dollar, aufgetragen über die letzten 100 Jahre und extrapoliert in die Zukunft. Zum Vergleich ist die Rechenleistung von Nervensystemen verschiedener Organismen angeführt (aus „Hans Moravec - Computer übernehmen die Macht“).

Auf Anwendungsebene profitiert die Molekulare Bioinformatik bereits jetzt von dem, was wir biologischen Systemen bislang abschauen konnten. Im Nervensystem müssen enorme Datenmengen intelligent verarbeitet werden. Entsprechend bieten sich künstliche neuronale Netze für die Verarbeitung der enormen Datenflut in der Biotechnologie an. Als künstliche neuronale Netze bezeichnet man im Computer simulierte Netzwerke schematisierter Neurone, mit denen sich an die Funktionsprinzipien des Nervensystems angelehnte Algorithmen realisieren lassen. Künstliche neuronale Netze werden bereits zur Klassifikation von DNA- und Proteinsequenzen eingesetzt, zur Vorhersage von Proteinstrukturen- und -funktionen, zur Auswertung von Genchipdaten oder demnächst zur datengetriebenen Modellierung von Genotyp-Phänotyp-Abbildungen basierend auf SNPs (Single Nucleotide Polymorphism), um nur einige Anwendungen zu nennen. Evolutionäre Algorithmen, Verfahren, die ähnlich wie die Evolution durch Mutation und Selektion Strukturen optimieren, werden z. B. zur Generierung alternativer Proteinsequenzen mit ähnlichen Funktionseigenschaften eingesetzt.

Was sind die großen Herausforderungen der Bioinformatik, was macht dieses Gebiet so spannend? Neben den oben genannten Themen, die für sich gesehen bereits Herausforderungen von hoher Relevanz darstellen, gibt es eine zentrale Frage, die dieses Gebiet treibt und von deren Antwort viele Anwendungsgebiete in hohem Maße profitieren würden: Warum sind biologische Informationsverarbeitungssysteme, insbesondere unser Nervensystem, heutigen Computern in entscheidenden Aspekten immer noch so haushoch überlegen? Diese Frage stellt sich immer drängender, da die Leistungsfähigkeit der Computer, gerechnet in Instruktionen pro Zeiteinheit, Jahr für Jahr rasant ansteigt. Abbildung 1 zeigt über die letzten hundert Jahre die Rechengeschwindigkeit, die man pro 1000 Dollar bekommen konnte. Diese doppeltlogarithmische Darstellung verdeutlicht das bekannte Moore'sche Gesetz, wonach sich die Rechengeschwindigkeit grob alle drei Jahre vervierfacht. Diese Darstellung extrapoliert diesen Trend, wofür es gute Gründe gibt, und zeigt zusätzlich noch, wie sich die jeweiligen Computer in ihrer Leistungsfähigkeit mit entsprechenden biologischen Organismen vergleichen. Dieser Vergleich kann natürlich nur eine Abschätzung sein, denn im Nervensystem kann man keine Rechengeschwindigkeit in Instruktionen/Zeiteinheit messen.

Es gibt verschiedene Ansätze, die Rechenleistung von Nervensystemen und Computern zu vergleichen. Der obige basiert darauf, dass man die visuellen Signalverarbeitungsschritte auf der Retina bereits sehr gut kennt und man weiß, wie viel Rechenleistung benötigt wird, diese Verarbeitungsschritte im Computer nachzubilden. Da die Retina ein ausgelagerter Bestandteil des Gehirns ist, kann dann über die Anzahl der involvierten Neurone auf Nervensysteme beliebiger Größe extrapoliert werden.

Diese Abschätzungen liefern als Resultat, dass bereits in rund 20 Jahren Computer prinzipiell an die Rechenleistung unseres Gehirns heranreichen werden. Bereits heute besitzen Computer die Rechenleistung des Nervensystems einer Eidechse. Jeder weiß, wie flink sich Eidechsen in ihrer Umwelt zurechtfinden. Wir sind nicht ansatzweise in der Lage, Roboter ähnlich intelligent in der Umwelt agieren zu lassen. Die Computerhardware wäre ausreichend leistungsfähig, jedoch sind wir nicht in der Lage, diese entsprechend zu programmieren. In 20 Jahren werden wir (ob wir es wollen, ist eine andere Frage) Computer bauen, die ähnlich leistungsfähig sein könnten, wie unser Gehirn. Uns fehlt es jedoch an den notwendigen Informationsverarbeitungskonzepten. Hier können wir noch viel von biologischen Systemen lernen. Die große Herausforderung ist, dass wir wissen, dass es geht. Die Natur macht es uns vor.

Abbildung 2 zeigt einen weiteren interessanten Zusammenhang. Es zeigt sich, dass Rechenleistung und Arbeitsspeicherkapazität in informationsverarbeitenden Systemen dazu tendieren, stets im gleichen Verhältnis zueinander zu stehen. Insbesondere ist interessant, dass dies nicht nur für technische, sondern auch für biologische Informationsverarbeitungssysteme gilt.

Welches zentrale Problem muss gelöst werden, um das Poteztial heutiger und zukünftiger Computer-Hardware in ähnlicher Weise zu nutzen, wie es uns biologische Systeme mit ihrer „Hardware" vormachen? Ein wesentliches Merkmal, welches wir biologischen Systemen abschauen möchten, ist deren Fähigkeit, sich auf verschiedenen Ebenen permanent an Umweltgegebenheiten anzupassen. Auf der einen oder anderen Ebene nennen wir es lernen. Diese Adaptivität ermöglicht es biologischen Systemen, Strukturen zu realisieren, die intelligent und robust in ihrer Umwelt agieren und auf höchst komplexen Steuerungs- und Kontrollmechanismen basieren. Adaptivität findet man auf verschiedenen Ebenen, am bekanntesten sind die Ebenen der genetischen und neuronalen Adaption. Adaption findet darüber auf verschiedenen Zeitskalen sowie auf Populations- und Individualebene statt.

Die zentrale Struktur, sowohl in biologischen als auch in zu konzipierenden technischen Systemen, ist die des sogenannten „autonomen Agenten". Abbildung 3 veranschaulicht diese zentrale Struktur. Ein solcher autonomer Agent soll basierend auf Informationen über den Zustand der Welt bestimmte Aufgaben lösen. Die global zu lösende Aufgabe biologischer Realisierungen solch autonomer Agenten ist implizit

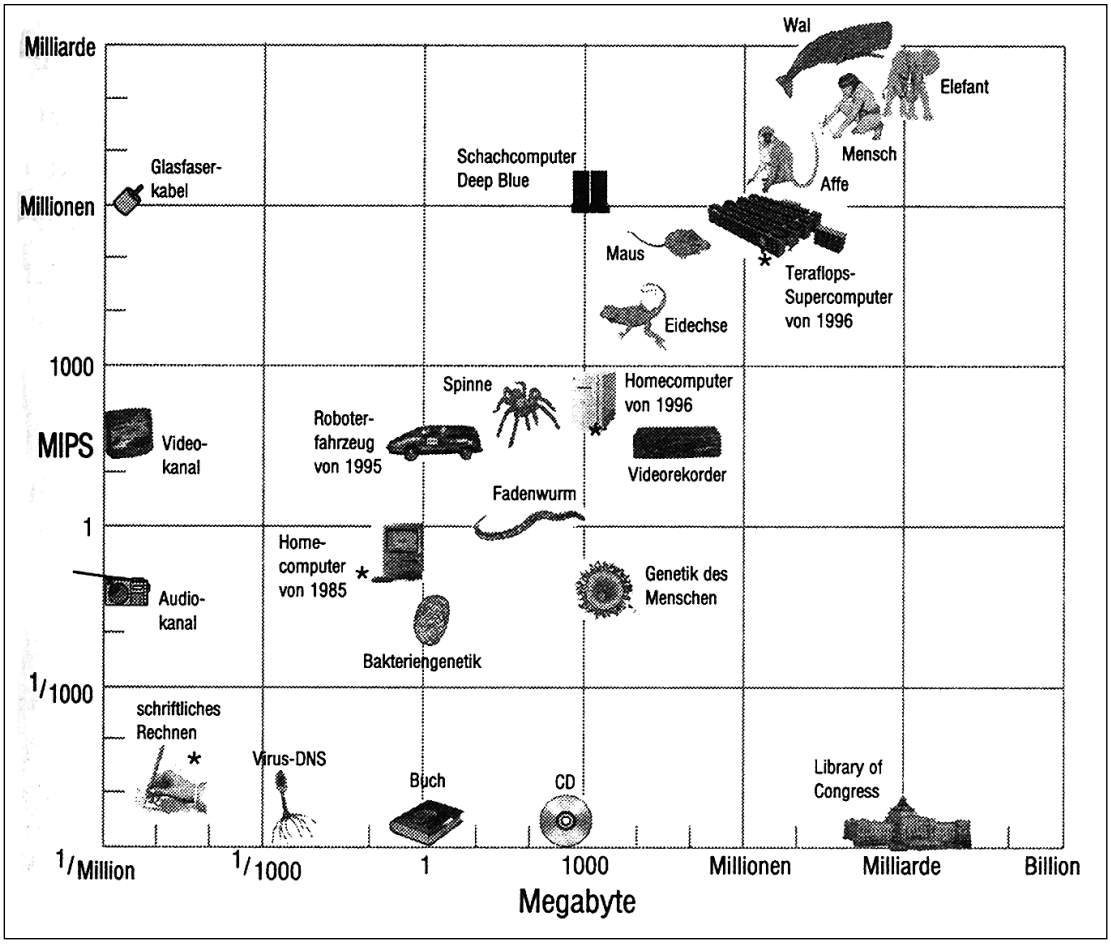

Abb. 2: Die Rechenleistung verschiedener technischer als auch biologischer Informationsverarbeitungssysteme in Millionen Instruktionen pro Sekunde (MIPS), aufgetragen gegen die in diesen Systemen jeweils verfügbare Arbeitsspeicherkapazität in Megabyte. Beide Systemgrößen stehen in etwa stets in einem eins-zu-eins Verhältnis. Entlang der x-Achse sind reine Informationsspeicher- und entlang der y-Achse reine Informationsübertragungssysteme aufgeführt (aus „Hans Moravec - Computer übernehmen die Macht").

gegeben, nämlich sich als Struktur zu erhalten. Biologische Agenten tun dies über einen cleveren Trick, über das Anlegen von Kopien (Fortpflanzung). In technischen Realisierungen wird die Aufgabe expliziter formuliert, etwa „sortiere Bauteile auf Fließband", „steuere Stahlbandwalzstraße möglichst kostengünstig", „finde gewünschte Information aus dem Internet" oder „bestimme den Phänotyp zum vorliegenden Genotyp". Wir wissen, dass manche dieser Aufgabenstellungen im Prinzip wesentlich besser als heutzutage von technischen Systemen gelöst werden könnten, denn ähnliche Aufgaben werden von biologischen autonomen Agenten als Teilaufgabe mit dem Ziel der Strukturerhaltung sehr erfolgreich bewältigt.

Um die gestellte Aufgabe erfolgreich zu lösen, muss der Agent zunächst die für die Bewältigung der Aufgabe notwendigen Informationen über den Zustand der Welt erhalten. Dies erfolgt über die für diese Informationen passende Sensorik. Darüber entsteht „seine Welt-

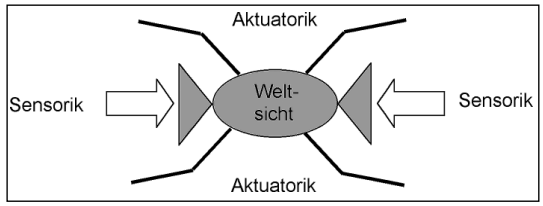

Abb. 3: Schematische Darstellung der prinzipiellen Struktur eines „autonomen Agenten".

sicht“, die über einen internen Zustand des Agenten repräsentiert wird. Ist der Weltzustand bezüglich der für die Aufgabenstellung wichtigen Aspekte erfasst, so muss der Agent die passende Aktion generieren. Dazu braucht er eine der Aufgabenstellung entsprechende Aktuatorik. Zum Beispiel muss eine Fledermaus zum Zwecke der Strukturerhaltung (überleben) Nahrung aufnehmen. Dazu sucht sie in vornehmlich dunklen Räumen umher. Informationen über die für diese Aufgabe wichtigen Aspekte wie räumliche Gegebenheiten und Aufenthaltsort von Beuteobjekten holt sie sich über ihre Ultraschallsensorik, um mit ihrem Fortbewegungs- und Fangapparat diese Informationen zur Lösung der Aufgabe umzusetzen. Entsprechend stattet man einen Roboter zum Beispiel mit einem Laserscanner aus, damit er darüber die Lage von Bauteilen auf einem Fließband erfassen kann, um diese dann mit seinem Greifer vom Band zu nehmen und zu sortieren.

Der autonome Agent muss also mit der passenden Sensorik, der passenden Aktuatorik und vor allem dem passenden Aktionsverhalten ausgestattet werden. Bei biologischen Systemen hat die Evolution alle diese drei Aspekte optimiert. Bei heutigen technischen Realisierungen wird die Sensorik und Aktuatorik vorgegeben. Auf diesen beiden Ebenen meint man in der Regel recht gut beurteilen zu können, was für die Lösung der gestellten Aufgabe notwendig ist. Wesentlich schwieriger ist es, den autonomen Agenten mit dem passenden Aktionsverhalten auszustatten. Häufig wird dieses fest vorgegeben. Basierend auf Modellen und Erfahrungswerten wird dem Agenten „falls Weltzustand=x, dann Aktion=y“ fest einprogrammiert. Selten ist jedoch eine zu einem gegebenen Weltzustand passende, geschweige denn optimale Aktion bekannt. Man hat es in der Regel mit hochdimensionalen Zustandräumen zu tun. Des Weiteren ist es häufig so, dass der nicht erfasste Teil der Welt sich ändert. Das kann dazu führen, dass die Aktionen, die ja nur auf dem erfassten Teil der Welt basieren, sich ebenfalls ändern, also adaptieren müssen. Die Gelenkmotoren eines Roboters können zum Beispiel über einen längeren Einsatz durch Verschleiß ihre Charakteristik verändern, weshalb das Steuerungssystem entsprechend seine Ansteuerung adaptieren muss.

Biologische Systeme sind hervorragend in der Lage, ihr Aktionsverhalten selbständig an die Gegebenheiten anzupassen. Vor allem darin liegt ihre deutliche Überlegenheit gegenüber technischen Systemen. Insofern liegt es nahe, sich die dahinterliegenden Prinzipien näher anzuschauen. Dies ist eines der Kernthemen der Bioinformatik und nicht nur von sehr hoher Praxis-, sondern auch Erkenntnisrelevanz. Denn wir kennen diese höchst erfolgreichen Informationsverarbeitungsprinzipien bislang nur ansatzweise.

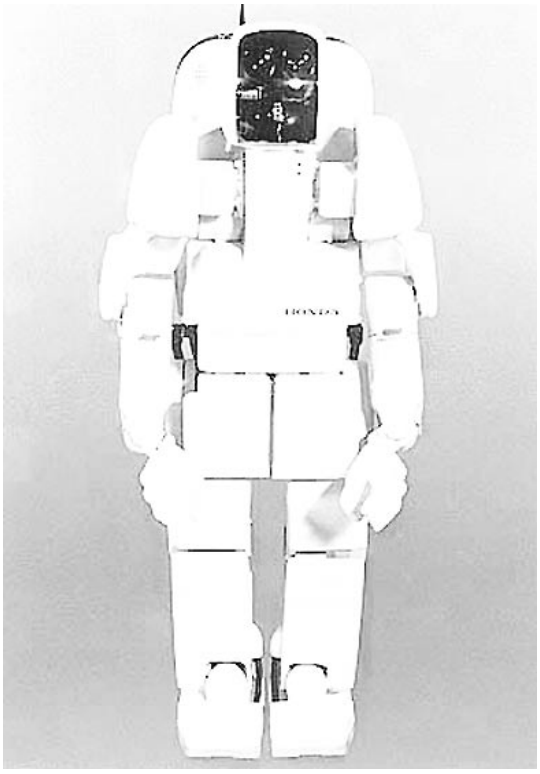

Abb. 4: Der derzeit komplexeste autonome Roboter (Agent) weltweit, der humanoide Roboter P3 von Honda. Er ist 1,60 m groß, 130 kg schwer und kann auf zwei Beinen sogar Treppen steigen.

Abstrakt formuliert ist es der Natur gelungen, ein Verfahren zu konzipieren, welches eine Abbildung von einem hochdimensionalen Eingaberaum auf einen Raum von Ausgabewerten anhand von einzelnen Stichproben lernen kann. Der Eingaberaum ist definiert durch all die Zustände, die die Welt des autonomen Agenten bezüglich seiner zu lösenden Aufgabe annehmen kann. Diese Zustände sind in der Regel durch viele Aspekte bestimmt, was diesen Raum sehr hochdimensional macht. Welche Aktion jetzt auf einen bestimmten „Weltzustand“ passt, soll das System lernen. Dazu muss es zunächst diejenigen Aspekte des Weltzustandes, die für ihn wichtig sind, durch seinen internen Zustand repräsentieren. Zum Beispiel muss die Lage des Objektes auf dem Fließband intern kodiert werden, jedoch nicht etwa welcher Arbeiter es auf das Fließband gelegt hat. Die interne Kodierung muss mächtig genug sein, um alle wichtigen Aspekte des Weltzustandes repräsentieren zu können. Dieses Problem kann man als gelöst ansehen, wenn man bedenkt, dass ein heutiger Computer in seinem Arbeitsspeicher im Prinzip $10^{100.000.000}$ verschiedene Zustände repräsentieren kann. Das menschliche Gehirn liegt allerdings noch weit

darüber. Auf der anderen Seite hat das gesamte Universum nicht mehr als 10^{100} Atome.

Das Problem liegt woanders. Wie kann dem System beigebracht werden, welche Aktion es jedem seiner internen Kodierungszustände zuordnen muss. Diese Aktion muss ja zu dem Weltzustand, der durch die jeweilige interne Kodierung repräsentiert wird, passen. Angesichts der überwältigend großen Zahl verschiedener Zustände kann dies unmöglich explizit geschehen. Auch die Natur hat dies nicht explizit tun können, erst recht nicht durch „trial and error", denn selbst das Erdalter wäre bei weitem nicht ausreichend gewesen.

Die Natur optimiert ihre Systeme zwar mit Hilfe von „trial and error", dies alleine genügt jedoch nicht. Entscheidend ist, dass der Agent von einem eventuell zufällig erlangten Lernerfolg verallgemeinern kann. Hat er durch eventuell zunächst zufällige Aktionen eine zum Weltzustand passende gefunden, so muss er in der Lage sein, daraus zu schließen, wie die Aktionen für ähnliche Weltzustände aussehen müssen. Dazu muss er „wissen", welche Zustände ähnlich sind in dem Sinne, dass sie ähnliche Aktionen erfordern. Die interne Kodierung der Weltzustände muss also diese Ähnlichkeiten erfassen. Nur dann ist es möglich, die im wahrsten Sinne überastronomische Zahl möglicher Zustände mit passenden Aktionen des Agenten zu belegen. Um sich als autonomer Agent, ob biologisch oder technisch realisiert, an die Gegebenheiten oder Änderungen „seiner Welt" adaptieren zu können, ist es also vor allem erforderlich, wie auch immer gewonnene Adaptionserfolge generalisieren zu können. Offensichtlich hat die Natur eine interne Kodierung zur Repräsentation der Weltzustände gefunden, die die Ähnlichkeitsstrukturen unserer Welt im Sinne der zu lösenden Aufgaben hervorragend erfasst und damit Generalisierungen ermöglicht. Dies gilt sowohl für den neuronalen als auch genetischen Kode.

Künstliche neuronale Netze versuchen, diese neuronale Kodierung zumindest ansatzweise zu imitieren, um damit Agenten zur Lösung technischer Aufgabestellungen zu konzipieren. Aber auch wenn der Agent mit einer guten Generalisierungsfähigkeit ausgestattet ist, so werden trotzdem immer noch viele Versuche benötigt, um bei komplexen Aufgabenstellungen die Gegebenheiten zu lernen. Eine Anwendung, wo künstliche neuronale Netze zum ersten Mal in größerem Maßstab kommerziell erfolgreich zum Einsatz kamen, ist bei der Steuerung von Stahlbandwalzstraßen. Abbildung 5 zeigt schematisiert solch eine Walzstraße. Ein auf rund 1000 Grad erhitzter Stahlblock von rund 10 m Länge läuft in die Walzstraße ein, und herauskommen soll ein rund 1 mm dickes und 1 km langes Stahlband, mit Dickentoleranzen im Mikrometerbereich. Verkauft werden diese Stahlbänder dann z. B. an Autohersteller. Die Steuerung solch einer Walzstraße ist eine hochkomplexe Angelegenheit, und die Frage war, ob ein Agent, ausgestattet mit künstlichen neuronalen Netzen, zumindest die Steuerung von Teilaspekten dieser Walzstraße lernen kann. Bei der Walzkraftsteuerung zum Beispiel bekommt der Agent als Information über den „Weltzustand" die Bandtemperatur, Walzgeschwindigkeit, Bandzüge, Walzenradien und Stahlsorte, und muss daraus als „Aktion" die für die gewünschte Enddicke notwendigen Walzkräfte an die Walzwerksteuerung weitergeben. Es braucht mehrere tausend Versuche, aber dann hat der Agent gelernt, welche Walzkräfte passen. Die erzielten Resultate übertreffen die mit konventionellen Techniken erreichte Qualität. Mit mehreren tausend Versuchen hat man bei weitem nicht alle möglichen Gegebenheiten, die auftreten können, abgedeckt. Aber offensichtlich ist der Agent mit seinen künstlichen neuronalen Netzen in der Lage, bei dieser Aufgabe gut zu generalisieren und z. B. vorher noch nicht gewalzte Stahlsorten unter vorher nicht angetroffenen Bedingungen gut zu walzen. Die von den künstlichen neuronalen Netzen gewählte Kodierung der „Weltzustände" hat offensichtlich deren Ähnlichkeiten im Sinne der Aufgabenstellung gut repräsentiert. Diese Agenten sind mittlerweile weltweit in Stahlproduktionsanlagen im Einsatz. Deren Erfolg hat dazu geführt, dass nun alle möglichen sensorischen Signale an solch einer Walzstraße aufgezeichnet werden. Den Agenten möchte man so viele „Weltzustände" wie möglich präsentieren, an denen sie sich

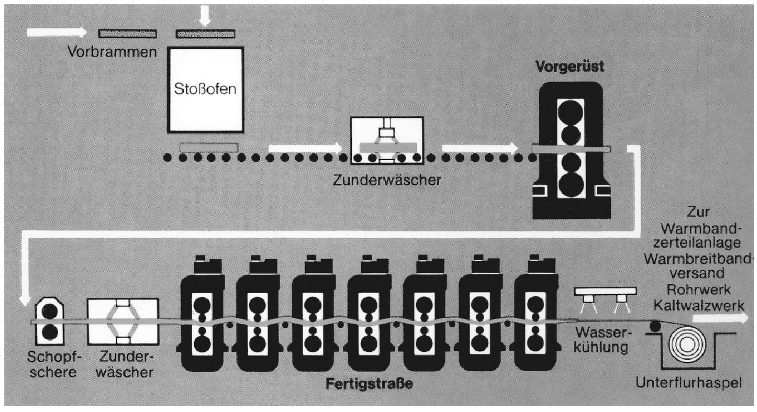

Abb. 5: Schematische Darstellung einer Walzstraße. Entscheidend ist die sogenannte Fertigstraße, wo das Stahlband auf Mikrometer genau auf wenige Millimeter Dicke gewalzt werden muss.

trainieren und „ihre Welt kennen lernen“ können. Die enorme Entwicklung der Computertechnologie (obiges Moore’sche Gesetz) wird es in einigen Jahren ermöglichen, die an solch einer Walzstraße anfallenden Daten über dessen gesamte Lebensdauer aufzuzeichnen.

Mit der Bioinformatik sind wir nun beim Stahlwerk gelandet. Das verdeutlicht die Breite dieses Gebietes. Ein kleiner Schritt bringt uns aber sofort wieder zurück zu den Anwendungsproblemen der molekularen Bioinformatik. Dieselben Konzepte, die dem Agenten erlauben, erfolgreich zur Steuerung eines Walzwerkes beizutragen, können ja vielleicht auch eingesetzt werden, um zum Beispiel Proteinfunktionen oder Genotyp-Phänotyp-Abbildungen zu lernen. Es liegen große Mengen von Daten vor, wobei die für eine bestimmte Problemstellung relevanten Daten allerdings schnell auf wenige zusammenschmelzen, zumindest im Vergleich zur Komplexität des Problems. Daher entscheidet auch hier wieder die Generalisierungsfähigkeit des Agenten.

Hochaktuell ist zum Beispiel der aller Ortens gestartete Versuch, über die Bestimmung sogenannter SNPs

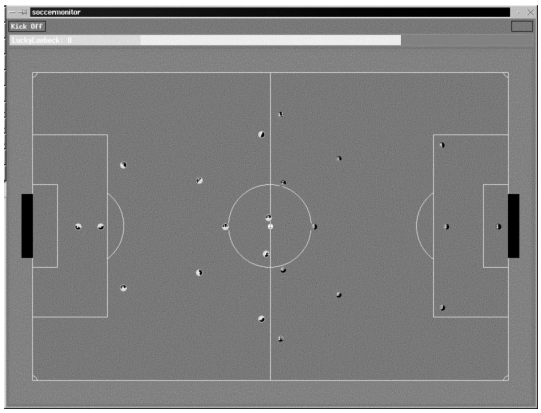

Abb. 6: Anstoß beim Spiel zweier RoboCup-Mannschaften der Softwareliga bei der Europameisterschaft 2000 in Amsterdam. Unter http://www.inb.mu-luebeck.de/robocup/robocup-d.html kann man sich die entscheidenden Szenen dieses Spiels anschauen (Lucky Lübeck gegen PSI aus Russland).

(Single Nucleotide Polymorphism) im Genom phänotypische Aspekte eines Organismus vorherzusagen. Die Daten der Sequenzierung des menschlichen Genoms führen zu der Schätzung, dass im menschlichen Genom im Mittel nur an jeder tausendsten Position eine Mutation aufgetreten ist. Könnte man die Genome aller Menschen übereinander legen, so würde man feststellen, das im Mittel nur an jeder tausendsten Stelle unterschiedliche Nukleotide auftreten und damit ein sogenannter Polymorphismus vorliegt. An 99.9 % der rund 3 Milliarden Positionen im menschlichen Genom liegt immer das gleiche Nukleotid vor. An drei Millionen Positionen können also Unterschiede auftreten. Geht man davon aus, dass nur rund drei Prozent des Genoms kodierend genutzt werden und rund 50.000 Gene vorliegen, so gibt es im Mittel auf jedem Gen nur rund 2 Positionen, auf denen dieses zwischen zwei Menschen variieren kann (es gibt Ausnahmen, Gene, die in einer wesentlich höheren Anzahl von Varianten vorkommen können). Da es praktisch ausgeschlossen ist, dass an derselben Position zweimal eine Mutation aufgetreten ist, werden dort höchstens zwei verschiedene Nukleotide auftreten können. Der Informationsgehalt eines menschlichen Gens beträgt also im Mittel lediglich 2 Bit und der relevante Teil des gesamten menschlichen Genoms damit nur 100.000 Bit (basierend auf dem Vorwissen, um welches Gen bzw. Genom es sich handelt. Natürlich sind dies Abschätzungen von Größenordnungen). Mit rund 13 Kilobyte kann das Genom eines menschlichen Individuums spezifiziert werden.

So wie beim Walzwerk ein Agent in der Lage ist, den Zusammenhang zwischen den aktuellen Gegebenheiten in der Walzstraße und den erforderlichen Walzkräften zu lernen, so wäre im Prinzip ein Agent vorstellbar, der basierend auf den 13 Kilobyte Informationen über den vorliegenden Genotyp den dazugehörigen Phänotyp vorhersagt. Davon sind wir natürlich weit entfernt, die Komplexität ist viel zu groß. Auf eingeschränkter Ebene möchte man dies aber versuchen. Lässt sich die Anzahl der für ein bestimmtes Krankheitsbild oder eine bestimmte Medikamentenverträglichkeit verantwortlichen Gene stark einschränken, so hat man vielleicht eine Chance. Die Information, auf der der Agent dann seine Entscheidung treffen muss, reduziert sich auf wenige Bit. Die „Welt“, in der der Agent sich zurechtfinden muss, könnte dann einfach genug werden. Entscheidend für den Erfolg wird es auch hier sein, dass dem Agenten ein Kodierungsschema für den „Weltzustand“ (Genotyp) mitgegeben wird, das die Ähnlichkeiten in dieser „Welt“ gut repräsentiert und damit dem Agenten Generalisierungen ermöglicht. Eine der großen Fragen der Bioinformatik wird es sein, dieses Kodierungsschema zu finden.

Von einzelnen Agenten ist es dann ein naheliegender Schritt zu Populationen von Agenten. Was gewinnt das Gesamtsystem, wenn jeder Agent nicht mehr nur für sich agiert, sondern die Agenten die Möglichkeit haben, miteinander zu kommunizieren, um eine Aufgabe gemeinsam zu lösen. Abhängig von der speziellen Wechselwirkung zwischen den Agenten ergibt sich ein spezielles globales Verhalten der gesamten Population. Ein in diesem Zusammenhang häufig zitiertes Beispiel ist der Ameisenhaufen. Viele für sich „dumme“ Individuen bringen durch entsprechende Interaktion ein hohärentes Verhalten der Gesamtpopulation hervor, so

dass solch beeindruckende Leistungen wie das Bauen von Ameisenhügeln erbracht werden. Welche Wechselwirkungen verursachen welches globale Verhalten, und, weit schwieriger, welche Wechselwirkungen benötigt man für ein gewünschtes globales Verhalten. Mit solchen Populationen von Agenten lassen sich zum Beispiel auch gesellschaftliche Phänomene simulieren und besser verstehen, bis hin zu solch abstrakten Aspekten wie die Auswirkungen gesellschaftlicher Verhaltensnormen. In den Wirtschaftwissenschaften versucht man mit solchen Agentenpopulationen, Wirtschaftsysteme zu modellieren. Die Auswirkungen von makroökonomischen Eingriffen können darüber simuliert und besser abgeschätzt werden.

Durch die Kommunikation zwischen Agenten kommen neue Dimensionen der Komplexität ins Spiel. Nicht ohne Grund vermutet man, dass der größte Teil des menschlichen Gehirns für die soziale Interaktion benötigt wird, darin also die größte Intelligenzleistung des Menschen liegt. Auf jeden Fall sind Computer hier ganz besonders schlecht. Erfolgreiche soziale Interaktion erfordert das Einfühlen in die Situation des anderen, was wiederum ein „Bewusstsein" des eigenen Ichs und das des Gegenübers voraussetzt. Es lässt sich spekulieren, ob sich mit Sozialverhalten auch Bewusstsein entwickeln muss und umgekehrt. Je höher entwickelt das Sozialverhalten einer Spezies, desto ausgeprägter deren Bewußsein? Dann sollte sich auch die Spezies Mensch in noch weit höhere Bewusstseinszustände begeben können.

Zurück zu dem, was heutzutage bereits möglich ist. Vor gut einem Jahr ist das Institut für Neuro-und Bioinformatik der MUL mit einem Team von Studenten beim sogenannten RoboCup eingestiegen. RoboCup ist ein weltweites wissenschaftliches Projekt, dessen Ziel die Erforschung von Konzepten und Techniken für die Entwicklung von Systemen interagierender autonomer Agenten ist. Als Szenario wird bei RoboCup ein Fußballspiel verwendet, bei dem die Spieler durch künstliche Agenten (Software- oder Hardwareroboter) repräsentiert werden. Dieses Szenario stellt für alle Aspekte der Multiagentenentwicklung eine Herausforderung dar, von den limitierten Informationen, die den Agenten zur Verfügung stehen, der motorisch-sensorischen Koordination, der Kooperation zwischen Agenten eines Teams zur Lösung der gemeinsamen Aufgabe bis hin zur Bewältigung widriger Umstände (in Form der Agenten des gegnerischen Teams). Von besonderer Bedeutung ist dabei die Tatsache, daß sich unterschiedliche wissenschaftliche Konzepte einander im direkten Vergleich gegenüberstellen lassen. So finden im Rahmen des RoboCup-Projektes alljährlich Weltmeisterschaften und weitere Turniere statt. Abbildung 6 zeigt die Szene beim Anstoß zu einem Spiel zweier Teams aus der Softwareliga bei der Europameisterschaft 2000 in Amsterdam. Beide Teams bestehen aus 11 Agenten. Jeder Agent bekommt abhängig von seiner Position auf dem Feld und seiner Blickrichtung bestimmte Informationen über die aktuelle Spielsituation, die er dann in passende Aktionen umsetzen muß. Jeder Agent muß dabei seine Rolle innerhalb der Mannschaft einnehmen. Der Torwart braucht ein anderes Aktionsverhalten als die Stürmer. Die ersten Spiele erinnerten an die der F-Jugend, an einen Schwarm von Spielern, der sich mit dem Ball bewegt. Unter http://www.inb.mu-luebeck.de/robocup/robocup-d.html kann man sich anschauen, dass sich seitdem viel getan hat. Trotzdem glauben wir natürlich nicht, dass sich bereits Bewusstsein bei den Agenten eingestellt hat.

Prof. Dr. rer. nat. Thomas Martinetz ist Direktor des Instituts für Neuro- und Bioinformatik der Medizinischen Universität zu Lübeck. Dem vorliegenden Beitrag liegt seine Antrittsvorlesung vor der Technisch-Naturwissenschaftlichen Fakultät vom 6. Juli 2000 zugrunde.