Informationstechnik und Bioinformatik in der Pflanzenzüchtung

WERNER SCHULTZ, EINBECK

Abstract

Information Techonology and Bioinformatics in Plant Breeding Present-day plant breeding is not only characterized by the close connection between traditional and modern plant-breeding techniques, but more and more by the increasing use of IT-systems resp. bioinformaticsystems.
In this field, many specific tools have been developed to generate and handle the rapidly increasing amount of available data, like plant breeding data bases, analyze-tools, systems for laboratory information management (LIMS), Mobile Computing, Biocomputing systems, Digital image processing etc..
A brief overview over some bioinformatic applications of KWS, one of the world´s leading plantbreeding companies, illustrates the benefits, but it also demonstrates, that certain limitations, mostly arising from the increasing complexity of the systems, have to be taken into consideration.

1 Einleitung

Das Bild der modernen Pflanzenzüchtung wird zunehmend durch den Einsatz der Informationstechnologie geprägt. Die vielfältigen Möglichkeiten der Informationsspeicherung und -auswertung nehmen heute einen herausragenden Stellenwert im gesamten Züchtungsumfeld von der Versuchsplanung über die Selektion bis hin zur Erfassung, Auswertung und Abspeicherung von Qualitätsdaten ein. Dazu liefert die Genom- und Züchtungsforschung molekulargenetische Methoden, die dem Züchter völlig neue Informationen eröffnen. Flankiert werden die verschiedenen spezifischen Anwendungen durch Systeme zum Management der im Labor gewonnenen Qualitätsdaten, durch Methoden der Digitalen Bildverarbeitung, durch Mobile Datenerfassungsgeräte (MDEs) etc. Hinzu kommen Kommunikationsmedien wie Internet / Intranet, Mailsysteme für den weltweiten Datenaustausch, Office-Anwendungen etc., deren Funktionalität heute als nahezu selbstverständlich vorausgesetzt wird - dabei leisteten sämtliche Anwendungen noch vor wenigen Jahren einen Bruchteil dessen, wozu sie heute in der Lage sind.
Dieser Gesamtkomplex der züchterischen Informationsverarbeitung wird im Forschungsumfeld der KWS unter dem Begriff der *Bioinformatik* zusammengefaßt. Diese Begriffsdefinition geht damit deutlich über die „klassische“ Auslegung hinaus, die i.d.R. den Focus auf die Molekularbiologie setzt. Vereinfacht gesagt beinhaltet Bioinformatik das Sammeln, Analysieren, Speichern und Nutzen biologischer Informationen zu züchterischen Zwecken. Ziel ist die durchgängige datentechnische Unterstützung der Zuchtprozesse vom Feldversuchswesen bis zur Biotechnologie. Wie vielschichtig dieses Gebiet tatsächlich ist, wird im folgenden anhand ausgewählter „Bausteine“ der Bioinformatik dargestellt.

2 Zuchtdatenmanagement am Beispiel der Zuckerrübe

Der prinzipielle Ablauf des Zuckerrüben-Züchtungsprozesses erscheint auf den ersten Blick überschaubar: Aus vorhandenem Züchtungssaatgut werden nach bestimmten Kriterien Kreuzungspartner ausgewählt, die in Feldversuchen angebaut werden. Mittels Bonituren und labortechnischer Untersuchungen wird die Qualität jeder Generation ermittelt, und zwar sowohl die des Saatgutes (Keimfähigkeit, Gewicht etc.) als auch die der daraus produzierten Pflanzen bzw. Rüben (z.B. Zuckergehalt, Resistenzeigenschaften, Schossverhalten etc.). Die gewonne-

nen Daten fließen wieder ein in die nachfolgende Versuchsplanung. Am Ende eines solchen Prozesses steht idealerweise die Produktion reinerbiger Vater- und Mutterlinien, aus denen schließlich Hybridsaatgut für den Verkauf in großen Mengen produziert wird.
Dieses Prinzip stellt jedoch einen Gesamtprozess dar, der in der Realität derart komplex ist, dass eine Abwicklung ohne Datenverarbeitung völlig undenkbar wäre. Die Vielzahl von Züchtungszielen und -methoden, Qualitätsuntersuchungen, Versuchsstandorten und -varianten resultiert in einem sehr hohen Datenvolumen, das nur mit einem hoch angepassten, leistungsfähigen Datenbanksystem zu bearbeiten ist. Folgende Faktoren sind entscheidend für die Komplexität:

- Allein für die Zuckerrübenzüchtung werden Informationen von mehreren hunderttausend Kreuzungsnachkommen zusammengeführt.
- Sämtliche Informationen zu einem bestimmten Saatgutposten über Abstammungen, Qualitäten, Standorte, Versuchsdesign etc. müssen dem Züchter zugänglich sein. Hierbei ist auch zu berücksichtigen, auf welche Art und Weise Elternpflanzen entstanden sind (aus Saatgut, aus Pflanzenteilen etc.).
- Wesentlich für die optimale Selektion ist die Tatsache, dass auch die Qualitätsinformationen aller Vorgängergenerationen zur Verfügung stehen.
- Für die Versuchsplanung müssen Informationen darüber einbezogen werden, welche Versuchsflächen an welchen Standorten zur Verfügung stehen. Diese für das Versuchsdesign erforderlichen Daten werden in anderen Systemen vorgehalten und via Schnittstelle abgeglichen.
- Mit einer Schnittstelle zum Lagerverwaltungssystem wird geprüft, ob für geplante Versuche überhaupt genügend Versuchssaatgut vorhanden ist (Bestandsführung). Bei nicht ausreichender Saatgutmenge müssen die im System als Plangrößen angelegten Versuche schnell angepasst werden können.
- Sämtliche im Verlauf eines Jahres anfallenden Arbeiten müssen im System abgebildet sein. Hierzu gehört zum Beispiel die Einspielung von Boniturdaten, die i.d.R. mittels Mobiler Datenerfassungsgeräte erfasst werden.
- Bei bestimmten Züchtungsstufen besteht nur wenig Zeit zwischen Anlieferung des Saatguts, Aufbereitung, Einlagerung und Einsatz im folgenden Zuchtprogramm.
- Es müssen unterschiedlichste Druckjobs angesteuert werden (Listen für die Entnahme aus dem Lager, Beschriftung von Tüten für das Saatgut, Drill-Listen für die Aussaat auf dem Feld, Etiketten für Kreuzungen, Boniturlisten, Ergebnislisten etc.).
- Die statistische Auswertung der Versuchsdaten (Varianzanalysen, Mittelwertberechnungen etc.) erfolgt ebenfalls in speziellen Software-Modulen (z.B. Plabstat). Auch hier wird der Datentransfer über entsprechende Schnittstellen gewährleistet.
- Zudem tritt das Datenbanksystem an die Stelle des Zuchtbuches, mit dem der Zuchtprozess für öffentliche Prüfstellen transparent gemacht wird.

Den Kern dieses Gesamtsystems - eine Integrierte Zuchtdatenbank - hat KWS in den vergangenen Jahren für die Zuckerrübenzüchtung neu aufgebaut. Das neue System wurde individuell entwickelt, da es auf dem Markt keine Software gibt, mit der die komplexen Züchtungsprozesse und die spezifischen Anforderungen der Züchter abgebildet werden können. Eine besondere Herausforderung aus IT-Sicht ist die durch die Komplexität beeinflusste Performance eines solchen Systems: Bei Abfragen müssen die angeforderten Daten über mehrere Generationen konsequent auf Konsistenz geprüft werden. Hierfür ist eine Vielzahl von Plausibilitätsprüfungen erforderlich, die idealerweise ohne nennenswerte Antwortzeiten im Hintergrund abgearbeitet werden sollten.
Vor allem muss die Zuchtdatenbank für einen schnellen, bilateralen Datentransfer mit anderen Systemen ausgelegt sein (Abb. 1). So muss z.B. Saatgut, bevor es im Feldversuch eingesetzt wird, nach bestimmten Verfahren aufbereitet werden, anschließend ist der Versand aus dem

Lager an die jeweiligen Züchtungsstandorte abzuwickeln. Dazu sind Bestände zu prüfen und Aufträge verschiedener Züchter zu koordinieren. Abhängig von Parametern wie Zuchtstufe, Aufbereitungsart, Standort etc. können Saatgutposten desselben Genotyps unterschiedliche Identifizierungsnummern zugeordnet werden. Die für die verschiedenen Aufgaben eingesetzten Applikationen werden über Schnittstellen abgeglichen, so dass dem jeweiligen Bearbeiter alle relevanten Informationen zu einem Saatgutposten zur Verfügung stehen, unabhängig davon, in welchem System die entsprechenden Daten gehalten werden.
Letztendlich kann erst durch eine sinnvolle Vernetzung der Daten führenden Systeme gewährleistet werden, dass dem Züchter sämtliche relevanten Informationen zur Verfügung stehen und der Züchtungsprozeß so effektiv wie möglich abgewickelt wird.

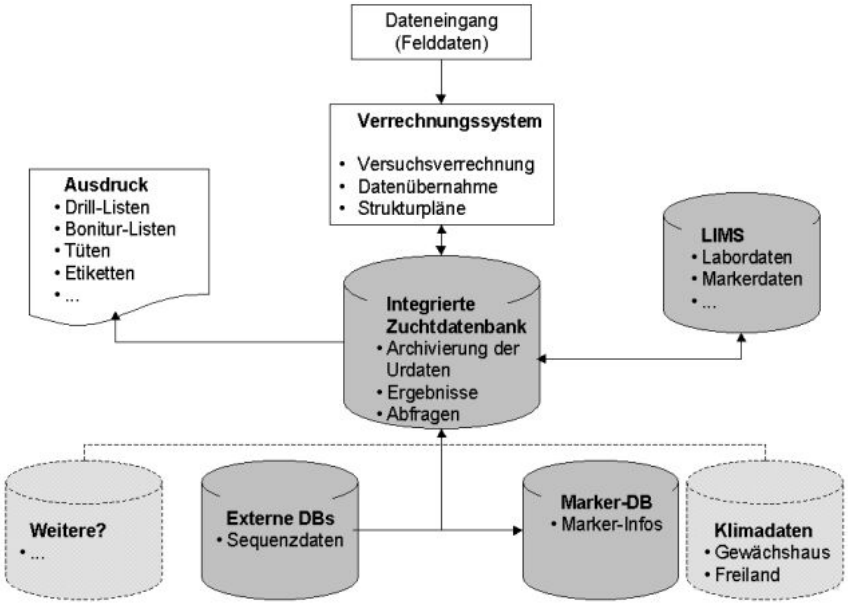

Abbildung 1: Wesentliche Bausteine der Zuchtdatenverwaltung Zuckerrübe

3 Labor- Informationsmanagement

Labortechnische Untersuchungen sind sowohl für das Qualitätsmanagement bei der Produktion von Verkaufssaatgut als auch im Züchtungsprozess von überragender Bedeutung. In Qualitätsuntersuchungen werden z.B. Saatguteigenschaften wie Keimfähigkeit, 1000-Korn-Masse, Einkeimigkeit, Vitalität usw. untersucht - viele Prüfungen sind von den amtlichen Sortenzulassungsstellen (Bundessortenamt) ohnehin vorgeschrieben. Mittels chemischer Analysen (HPLC, Elektrophorese etc.) werden Inhaltsstoffe, Hüllmassenqualität, Homozygotie, Ploidiestufe etc. untersucht.
Auch dieser Bereich der Pflanzenzüchtung ist durch ein stetig wachsendes Mengengerüst gekennzeichnet. Um einen hohen Durchsatz an Untersuchungen bei gleichbleibender Qualität gewährleisten zu können, ist es erforderlich, die Laborprozesse durch IT-gestütztes Datenmanagement zu unterlegen. Entsprechende, von verschiedenen Herstellern angebotene Labor-Informations-Management-Systeme (LIMS) beinhalten meist zahlreiche Standard-Features, müssen aber immer an die spezifischen Laborbedingungen angepasst werden. Sind die Prozesse erst abgebildet und die Laborgeräte integriert, bietet ein LIMS entscheidende Vorteile:

- Erfassung und Ablage aller Labordaten in einer Datenbank
- Gewährleistung einer lückenlosen Probenverfolgung

- Geregelte Zugriffsrechte für alle Nutzer
- Höherer Durchsatz durch Zeitersparnis bei der Datenerfassung
- Möglichkeiten der Datensichtung / -auswertung
- Elektronische Schnittstellen zur Auftrags- und Datenerfassung
- Qualitätsverbesserung durch Sicherheit vor Übertragungsfehlern
- Integriertes Bestellsystem für Untersuchungsmaterialien

Das Prinzip eines solchen LIMS-gestützen Ablaufs veranschaulicht **Abbildung 1**: Werden Qualitätsuntersuchungen unternehmensintern in Auftrag gegeben, so erfolgt dies ausschliesslich über das LIMS. Im System sind die für die verschiedenen Untersuchungen erforderlichen Arbeitsschritte hinterlegt. Parallel zur Auftragserfassung wird die zu untersuchende Probe zum Labor geschickt. Der gesamte Prozess erfolgt Barcode-gestützt, so dass die Informationen der jeweiligen Probe eindeutig zugeordnet werden können. Die Analysegeräte im Labor sind ebenfalls in Gesamt-Prozess integriert: Jede Analysestation ist mit einem Datenerfassungsgerät ausgestattet. Vor Untersuchung einer Probe beschränkt sich die administrative Arbeit auf das Einscannen des Barcodes an der Probe. Die gewonnenen Ergebnisdaten werden über Schnittstellen zurück in das LIMS übertragen. Hier können sie dann vom Auftraggeber abgerufen werden. Mit dem System kann zudem lückenlos verfolgt werden, welche Arbeitsstationen eine Probe bereits passiert hat.
Auch in anderen Bereichen mit intensivem Einsatz von Laborgeräten, z.B. in der Gewebekultur, wird LIMS zur Unterstützung der Arbeitsabläufe eingesetzt.
Durch die Optimierung der Datenflüsse bei Labor-Arbeitsabläufen und die sofortige Verfügbarkeit der Ergebnisse hat das bei KWS eingesetzte LIMS Nautilus von Thermolab Systems erheblich zur Optimierung des Züchtungsprozesses beigetragen und wird durch sukzessive Implementierung von Schnittstellen an weitere Systeme immer mehr an Bedeutung gewinnen.

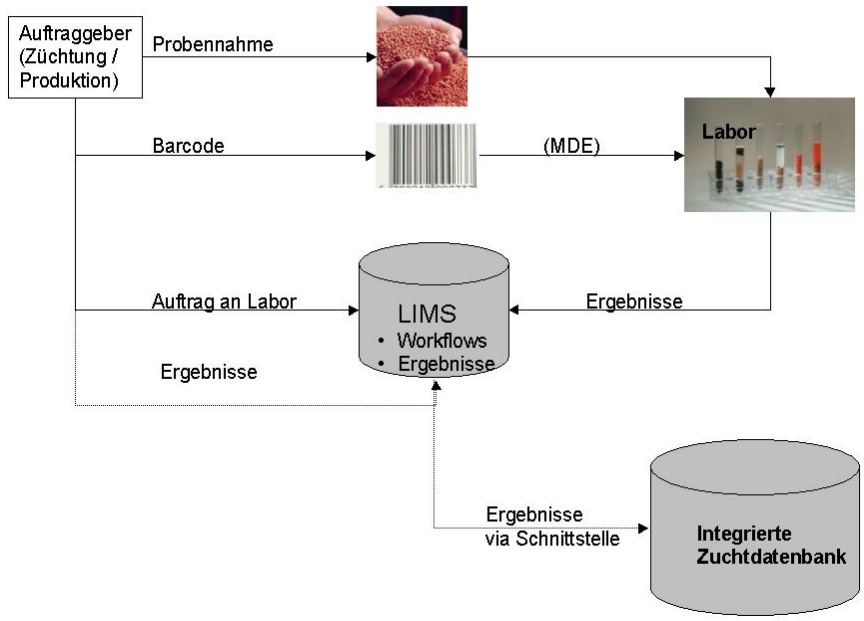

Abbildung 1: Prinzip des Labor-Informations-Management-Systems zur Unterstützung der Probenanalyse

4 Biocomputing und Marker-Technologie

Das Biocomputing, d.h. die informatorische Verarbeitung von DNA-Sequenzdaten und die Markerentwicklung und -nutzung, entwickelt sich zu einem immer wichtigeren züchterischen Werkzeug. Marker - so die allgemeine Definition - ermöglichen es, beliebige Individuen einer Population exakt zu beschreiben und damit von anderen Individuen abzugrenzen. Das Ziel: die Untersuchung von Pflanzen auf molekularer Ebene dahingehend, ob bestimmte Eigenschaften (z.B. Krankheitsresistenz) vererbt wurden. Ein wesentlicher Vorteil: Jungpflanzen können in einem sehr frühen Stadium auf das Vorhandensein von Merkmalen geprüft werden. Dies kann im gesamten Züchtungsablauf zu erheblicher Zeiteinsparung und zur Optimierung des Züchtungserfolges führen.
Zur Markerentwicklung werden zunächst im Labor extrahierte Sequenzen von DNA oder Aminosäuren in öffentlichen oder lizenzierten Datenbanken mit bereits bekannten Sequenzen verglichen. Wird festgestellt, dass eine Sequenz in anderen Organismen mit bestimmten züchterisch relevanten Eigenschaften korreliert, so schließen sich weitere Untersuchungen an, in denen überprüft wird, ob sich dieses „Kandidatengen“ auch bei der untersuchten Kulturpflanze zu einem Marker weiterentwickeln lässt, der schließlich in der Züchtung eingesetzt werden kann (**Abbildung 2**, links). Hierbei kommen neben molekularbiologischen Verfahren auch komplexe statistische Untersuchungen zum Einsatz. Ein wesentlicher Schwerpunkt ist die abschließende Verifizierung in Feldversuchen. Hier werden die Genotyp-Phänotyp-Interaktionen geprüft, d.h., die Pflanzen aus der spaltenden Nachfolgegeneration werden zunächst phänotypisch auf die gesuchten Eigenschaften überprüft, anschließend wird untersucht, ob der jeweilige Phänotyp mit einem Marker eng korreliert bzw. gekoppelt ist. Die so gewonnenen Markerdaten bzw. die daraus abzuleitenden züchtungsrelevanten Informationen werden schließlich in speziellen Datenbanken abgelegt, um sie den Züchtern zugänglich zu machen.
Um die Vielzahl der öffentlichen oder lizenzierten Datenbanken, in denen Sequenzdaten archiviert sind, optimal nutzen zu können, bieten diverse Unternehmen wie Biomax (München), Lion Bioscience (Heidelberg) oder Informax (US) spezielle Dienstleistungen an. Diese beinhalten in der Regel das Betreiben bzw. Aggregieren der Datenbanken sowie die laufende Weiterentwicklung der zugehörigen Analysesoftware, denn es bedarf effektiver Abfragealgorithmen, um die Informationen nutzen zu können. Züchtungsinstitutionen nutzen den Service, um z.B. mittels Ähnlichkeitsanalysen (sog. Homologievergleiche) in den öffentlichen Sequenzdatenbanken Informationen zu unbekannten Genen zu finden.
Dieser Bereich veranschaulicht die stetig wachsende Verzahnung von klassischen und modernen Züchtungsmethoden und den Einfluss der Informationstechnologie auf den Züchtungsfortschritt. Neben der Optimierung der labortechnischen Untersuchungen ist es vor allem die Datenverarbeitung, die eine Hochdurchsatz-Analyse überhaupt erst ermöglicht.

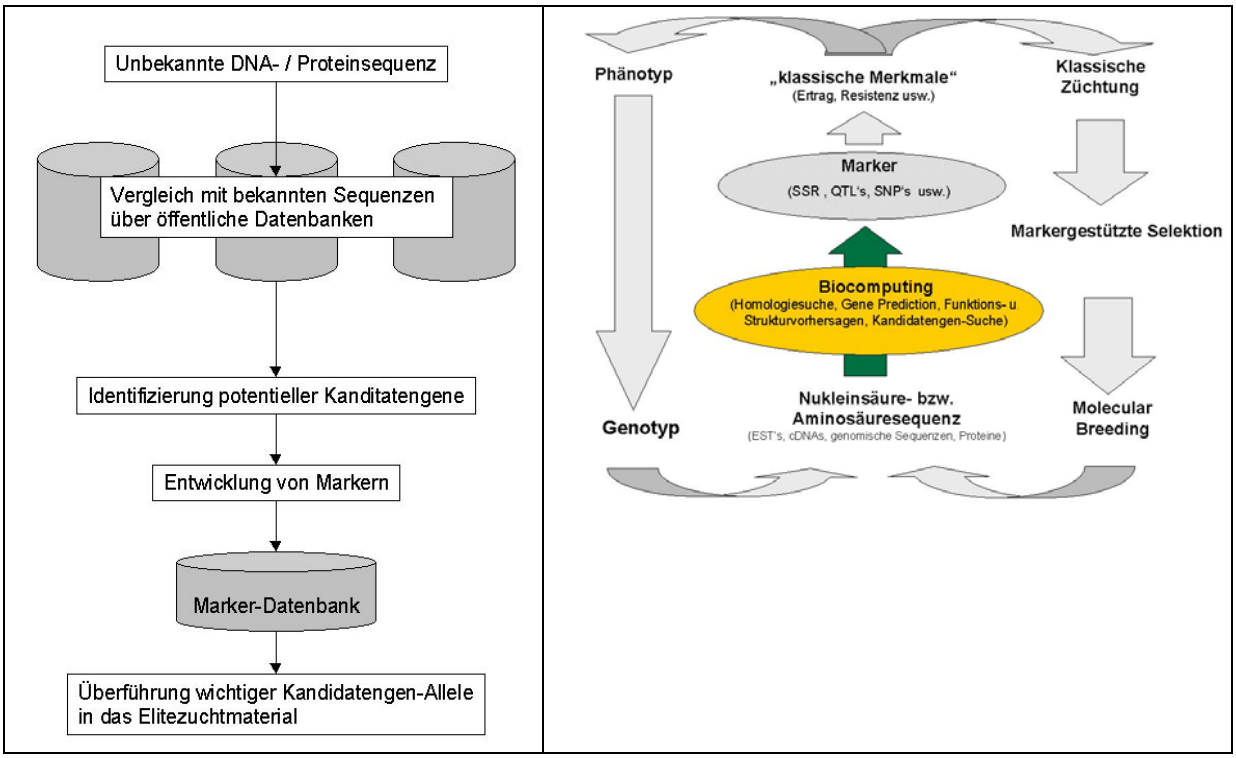

Abbildung 2: Prinzipieller Ablauf der Sequenzanalyse bis zur Markerentwicklung (links) und Einordnung von Sequenzanalyse und Markertechnologie in den Züchtungsprozess (rechts).

5 Weitere Bausteine der Bioinformatik

Neben der stetig wachsenden Menge an Informationen aus dem direkten Züchtungsumfeld steigt auch im F&E-Bereich die Fülle an Sekundärinformationen, auf die man möglichst effektiv zugreifen können muß. Hierbei bietet die moderne Informationsverarbeitung sehr günstige Voraussetzungen, wie an einigen Beispielen veranschaulicht werden kann:

- **Internet und Mail:** Mit der rasanten Ausbreitung des Internets und der steigenden Geschwindigkeit der Datenübertragung ist die Bedeutung des Internets für die Informationsbeschaffung erheblich gestiegen. Die Möglichkeit, Daten nahezu ortsunabhängig auszutauschen, führte dazu, dass sich das Internet innerhalb weniger Jahre zu einem der wesentlichen Recherchemedien etabliert hat. Zudem besteht die Möglichkeit des Informationsaustausches in Fachgruppen, die teilweise öffentlich zugänglich sind, teilweise geschlossenen Benutzergruppen vorbehalten sind. Auch die „klassische“ Literaturrecherche vereinfachte sich wesentlich, seit man via Internet Zugang zu öffentlichen Bibliotheken hat. Der Datenaustausch per Mail kann mittlerweile als selbstverständlich angesehen werden. Groupware-Anwendungen wie Lotus Notes bilden hier eine wesentliche Basis zur schnellen Informationsübermittlung innerhalb einer internen Arbeitsgruppe, zwischen verschiedenen Zuchtstationen oder mit externen Partnern.
- **Datenbanken für Informations- und Projektmanagement**: Bei der Abwicklung von Projekten wird - nahezu unabhängig von Art und Inhalt - eine Vielzahl von Informationen generiert, die idealerweise nicht redundant und für alle Beteiligten schnell verfügbar bereitgestellt werden sollen. Auch hier lassen sich Systeme wie Lotus-Notes sinnvoll einsetzen. Protokolle, To-Do-Listen, Projektpläne, beschreibende Dokumente, Tabellen, Bilder etc. werden dabei in einer zentralen Datenbank abgespeichert, auf welche die Beteiligten mit Zugangsberechtigung zugreifen können. Komfortable Datenbank-Optionen wie Volltextsuche, Filterfunktionen etc. erleichtern die Informationsverwaltung. Groupware-

Datenbanken lassen sich auch für weitere Anwendungsgebiete wie Literaturverwaltung, Patentverwaltung etc. sehr gut einsetzen.

- **Digitale Bildverarbeitung:** Auch in der Pflanzenzüchtung werden Methoden der Digitalen Bildverarbeitung vermehrt angewendet. Einsatzbereiche wie Feldaufgangszählungen im Freiland oder diverse Qualitätsprüfungen, in denen ähnliche Objekte ausgezählt oder auf das Vorhandensein bestimmter Objekte überprüft werden, bieten - wie in anderen Fachrichtungen auch - noch enormes Entwicklungspotential, während andere Anwendungen bereits etabliert sind. So können mittels Digitaler Bildverarbeitung im Rahmen von Qualitätsuntersuchungen Formparameter von Zuckerrüben bestimmt werden. Dazu werden die Rüben fotografiert, anschließend werden die Bilder mit einer individuell entwickelten Software ausgewertet. Die Ergebnisdaten (Kopfanteil, Länge etc.) werden dann wiederum in die Integrierte Zuchtdatenbank überführt, so dass sie den Züchtern zur Verfügung stehen. Welche Möglichkeiten sich hier langfristig bieten, ist gegenwärtig schwer abzuschätzen. Angesichts der stetig wachsenden Leistungsfähigkeit von Hard- und Software ist aber zu erwarten, dass dieser Bereich künftig weiter an Bedeutung gewinnen wird.
- **Gewächshaus- und Klimadatenmanagement**: Betrachtet man Bioinformatik im Sinne von datentechnischer Unterstützung der Zuchtprozesse, so lässt sich das Spektrum noch deutlich ausweiten, z.B. auf das Klimadatenmanagement: Um den Einfluß der Standortfaktoren auf die Pflanzenqualität bewerten zu können, wäre es ideal für den Züchter, wenn er zusätzlich zu den in der Zuchtdatenbank hinterlegten Qualitätsinformationen auch sämtliche Informationen zu den jeweiligen Standort- bzw. Klimabedingungen nutzen könnte. Die Meßverfahren für Einstrahlung, Luftzustand etc. sind etabliert, aber - bei entsprechender Meßqualität - teuer, vor allem aber zieht die Aufbereitung und Interpretation der gewonnenen Meßdaten einen nicht unerheblichen Arbeitsaufwand nach sich. Hinsichtlich der Kulturführung im Gewächshaus ist festzuhalten, daß moderne Klimaregelsysteme heute zwar eine sehr hohe Regelgüte aufweisen, sie sind aber häufig in Bezug auf die Auswertemöglichkeiten beschränkt. So muss das bei KWS eingesetzte Klimaregelsystem um eine Klimadatenbank ergänzt werden, in welche die generierten Klimadaten automatisch übertragen werden. Erst die datenbank-spezifischen Werkzeuge ermöglichen schließlich Auswertungen der Versuchsdaten in einem zufriedenstellenden Umfang. Mittelfristig ist hier zu prüfen, inwieweit eine Anbindung der Klimadatenbank an die Integrierte Zuchtdatenbank sinnvoll ist.

6 Ausblick

Bei Betrachtung der züchterischen Bioinformatik lassen sich deutliche Parallelen zu anderen IT-Einsatzbereichen ziehen: Einerseits steigt - dank IT-gestützter Verfahren - die Menge an verfügbaren, hochwertigen Informationen ständig und immer schneller an, und viele Arbeitsabläufe werden erheblich verbessert. Auch ist die Leistungsfähigkeit moderner Rechnersysteme längst kein begrenzender Faktor mehr, und durch eine weitestgehende Standardisierung von Schnittstellen kann der Datenaustausch zwischen verschiedenen Systemen erheblich vereinfacht werden.

Andererseits müssen für die Bearbeitung und Auswertung der gewonnenen Informationen immer wieder neue IT-Applikationen entwickelt werden, mit deren Umgang sich die Anwender vertraut machen müssen. Damit steht man bei der Entwicklung von Informationssystemen immer stärker (und immer häufiger) vor dem Konflikt, komplexe Prozesse möglichst umfassend in einem System abzubilden, das trotzdem für Anpassungen und Entwicklungen offen bleibt, fehlerunanfällig und unkompliziert zu handhaben ist. In der Bewältigung dieser Problematik liegt möglicherweise eine der wesentlichen Herausforderungen der züchterischen Bioinformatik.