Verwendung von Datenbankprogrammen in der Pflanzenzüchtung am Beispiel MS Access

M. HECKENBERGER, A.E. MELCHINGER und M. FRISCH

1. Einleitung

Der Zuchtfortschritt in der Pflanzenzüchtung basiert einerseits auf der schöpferischen Intuition der Züchter und dem positiven Verlauf vieler zufälliger Rekombinationsereignisse, andererseits aber auch zu einem nicht unerheblichen Teil aus der gezielten Ausnutzung von Informationen, die über die potentiellen Eltern von Züchtungspopulationen vorhanden sind. In der klassischen Pflanzenzüchtung werden zur gezielten Auswahl der Eltern hauptsächlich phänotypische und agronomische Merkmale herangezogen, die durch Messungen (z.B. Kornertrag, Wuchshöhe, TKG, Tage bis zur Blüte) oder Bonituren (z.B. Antherenfarbe, Blattstellung, Kornform, Resistenzmerkmale) erfasst werden.

In den letzten Jahren wurde die klassische Pflanzenzüchtung zunehmend durch molekularbiologische Analysen, wie z.B. Isoenzymbestimmung, molekulare Marker oder DNA Sequenzierung ergänzt. Alle Methoden resultieren in Informationen, die zusätzlich zu den standardmäßig erhobenen phänotypischen und agronomischen Daten zur Verfügung stehen. Während bei Methoden wie der Isoenzymanalyse oder der elektrophoretischen Untersuchung der Speicherproteine gezielte Informationen zum Proteinmuster der jeweiligen Genotypen erfasst werden, ermöglichen molekulare Marker und Sequenzanalysen einen direkten Blick ins Genom und somit auf den Genotyp potentieller Eltern und deren Nachkommen. Bei den mehr oder weniger zufällig im Genom vorkommenden molekularen Markern ist die Präsenz oder Absenz einer Markerbande die Grundlage für Diversitätsstudien oder der Suche nach Kandidatengenen für züchterisch und agronomisch relevante Merkmale. Sequenzinformationen helfen beispielsweise bei der Generierung von Primern auf der Suche nach neuen DNA-Markern, bei Genexpressionsstudien und letztlich auch bei der Isolierung von Genen.

Diese molekularen Daten werden gegenwärtig entweder mit Statistik-Softwarepaketen, wie SAS (SAS-Institute, 1998), R (IHAKA et al., 1996) oder S-PLUS (CHAMBERS, 1998) analysiert, oder aber mit Softwarepaketen, die eigens für diesen Zweck programmiert wurden. Dazu zählen beispielsweise NTSys (ROHLF, 1989), Arlequin (SCHNEIDER et al., 1997), G-MENDEL (HOLLOWAY et al., 1993), PLABQTL (UTZ und MELCHINGER, 1996), oder MAPMAKER (LANDER et al., 1987), die Antworten auf die verschiedensten züchterischen und molekularbiologischen Fragestellungen erlauben. Oftmals werden die entsprechenden Daten entweder als Inputfiles der jeweiligen Softwarepakete gespeichert, oder als Tabellenblätter in Tabellenkalkulationsprogrammen. Da keine dieser Lösungen speziell zur Datenspeicherung entwickelt wurde, ergeben sich daraus häufig die folgenden Probleme (nach FRISCH et al., 2002):

- Da die Inputformate der unterschiedlichen Softwarepakete sehr unterschiedlich sind, werden identische Daten häufig mehrmals in unterschiedlichen Formaten gespeichert. Müssen Daten geändert werden, führt dies oftmals zu Dateninkonsistenz, wenn nicht alle vorhandenen Datenfiles in gleichem Maße geändert werden. Außerdem führt diese Redundanz zu einem unnötig hohen Speicherbedarf.
- Nur der Versuchsansteller selbst kann die Codierung der Genotypen und die Datenstruktur reproduzieren. Dies führt zu Komplikationen bei der Reanalyse von Daten, hauptsächlich dann, wenn die Reanalyse lange Zeit nach der ersten Analyse erfolgt.
- Je nach Größe der Datensatzes wird bei der manuellen Umsortierung und Umgruppierung von Daten in das Inputformat eines anderen Softwarepaketes sehr viel Zeit verbraucht. Darüber hinaus birgt diese Methode ein großes Fehlerpotential.
- Es besteht keine Verbindung zwischen experimentell erhobenen und logistischen Daten, wie etwa Saatgutmanagement oder Laborbereich.
- Da die Datensätze meist nur für ein spezielles Experiment, mit einer individuellen Kombination aus Markern, Genotypen und Versuchsstandorten erstellt werden, ist die Kombination von Daten aus verschiedenen Experimenten zu einer gemeinsamen Analyse schwierig.

Mit der zunehmenden Markerdichte im Genom aller wichtigen Kulturarten, den immer schneller generierbaren Sequenzinformationen, sowie immer effizienteren Genexpressionsstudien kommt allerdings beim zunehmenden Zeitdruck und der teilweise unsicheren wirtschaftlichen Lage einer schnellen Verfügbarkeit von erhobenen Daten eine immer entscheidendere Bedeutung zu. Eine Voraussetzung dafür ist eine effiziente Datenspeicherung und -aufbereitung.

Neben der Studie von FRISCH et al. zur Speicherung von DNA-Markerdaten in Datenbanken gibt es zum gegenwärtigen Zeitpunkt nach unserem derzeitigen Wissensstand kein Konzept zur effizienten Speicherung aller in einem Zuchtprogramm anfallenden Daten. Gegenstand dieser Studie war die Entwicklung eines Datenbankkonzeptes zur verknüpften Speicherung von molekularen, agronomischen, morphologischen und logistischen Daten. Die hier vorgestellte Datenbankstruktur vermeidet Datenredundanzen, ermöglicht eine effiziente Speicherung von Daten in Standardformaten und er-

Autoren: Dipl.-Ing. sc. agr. Martin HECKENBERGER, Prof. Dr. Albrecht E. MELCHINGER und Dr. Matthias FRISCH, Institut für Pflanzenzüchtung, Saatgutforschung und Populationsgenetik, Universität Hohenheim, Fruwirthstr. 21, D-70593 STUTTGART

leichtert Eingabe, Aufbereitung, Austausch und Reanalyse von Daten.

2. Datenbankstruktur

Per Definition ist eine Datenbank ein "System zur Beschreibung, Speicherung und Aufbereitung von umfangreichen Datenmengen, die von verschiedenen Anwendungsprogrammen benutzt werden" (Duden Informatik, 2000). Datenbankprogramme sind in vielen verschiedenen Größen und Ausführungen auf dem Markt erhältlich. Während MS Access als Bestandteil von Microsoft Office oder DBase hauptsächlich für kleinere und einfach zu handhabende Datenbankanwendungen konzipiert wurden, wurden die Datenbankmanagementsyteme "DB2", "MS SQL Server", "Sybase" oder "Oracle" hauptsächlich für professionelle Anwendungen entwickelt und erfordern dementsprechende Investitionen und auch Expertise. Eine Alternative hierzu stellen die Open Source-Programme "PostgreSQL" oder "MySQL" dar.

Zur Wahrung einer hohen Flexibilität ist eine Grundlage von relationalen Datenbanken die möglichst feine Strukturierung aller Daten in unterschiedliche Tabellen. Feld- und Markerdaten sollten daher beispielsweise in getrennten Tabellen gespeichert und bei Bedarf wieder verknüpft werden. Auch innerhalb der Markerdaten sollten die generierten Daten getrennt von der Information über die Marker selbst oder den Daten über die Genotypen gespeichert werden.

Herzstück und zentraler Knotenpunkt des hier vorgestellten Datenbankmodells sind die Tabellen "list_of_genotypes" und "list_of_studies". Während in der ersteren alle Informationen über die im Zuchtprogramm verwendeten Genotypen, abgelegt sind, z.B. Pedigreeinformationen oder Informationen, zu welcher heterotischen Gruppe der Genotyp zuzuordnen ist, werden in der zweiten verschiedene Experimente definiert und voneinander abgegrenzt. Ausgehend von diesen Tabellen unterteilt sich die Datenbank in 4 Hauptmodule: molekulare Daten, Felddaten, Laborbereich und Logistik.

2.1 Modul: molekulare Daten

Kernstück des Moduls "molekulare Daten" ist die Tabelle "observed_marker_ data" (*Tabelle 1*), in der die Ergebnisse der durchgeführten Experimente abgelegt sind. Das Feld (=Spalte) "data_point_ID" steht für eine eindeutige Identifikationsnummer (=ID), anhand der jeder Datensatz eindeutig aufzufinden ist. Die Felder "study_ID", "marker_ID" und "genotype_ID" beziehen sich auf die gleichnamigen Felder in den verknüpften Tabellen (*Abbildung 1*). Im Feld "allele_name" werden die am jeweiligen Locus vorliegenden Allele definiert. Entweder über eine ID und eine zusätzliche Tabelle "list_of_alleles" oder z.B. über die Größe der DNA-Fragmente in Basenpaaren (bp) (*Tabelle 1*). Im Feld "allele_state" ist angegeben, ob für den entsprechenden Genotyp das Allel vorhanden (=1) oder nicht vorhanden (=0) ist, oder ob keine Informationen (=9) über den Genotyp für den jeweiligen Marker vorhanden sind, z.B. bei schlechter DNA-Qualität oder fehlender Amplifikation. In den meisten Fällen genügt es dabei, nur die Datensätze mit "1" oder "9" abzuspeichern.

Informationen über die verwendeten Marker, wie deren Position im Genom, die Art ihrer Vererbung oder Markertyp sind in der Tabelle "list_of_markers" (*Tabelle 2*) gespeichert, wobei beide Tabellen über die Spalte "marker_ID" miteinander verknüpft sind. "Verknüpfung" bedeutet in diesem Fall, dass sich die Identifikationsnummer "marker_ID" in

Tabelle 1: **Struktur der Tabelle "observed_marker_data"**

data_point_ID	marker_ID	allele_name	genotype_ID	allele_state
…	…	…	…	…
3	1078	200	1	9
4	1078	202	1	9
5	1079	181	1	0
6	1079	183	1	0
7	1079	185	1	1
8	1079	187	1	1
9	1079	189	1	0
10	1080	153	1	1
11	1080	155	1	0
…	…	…	…	..

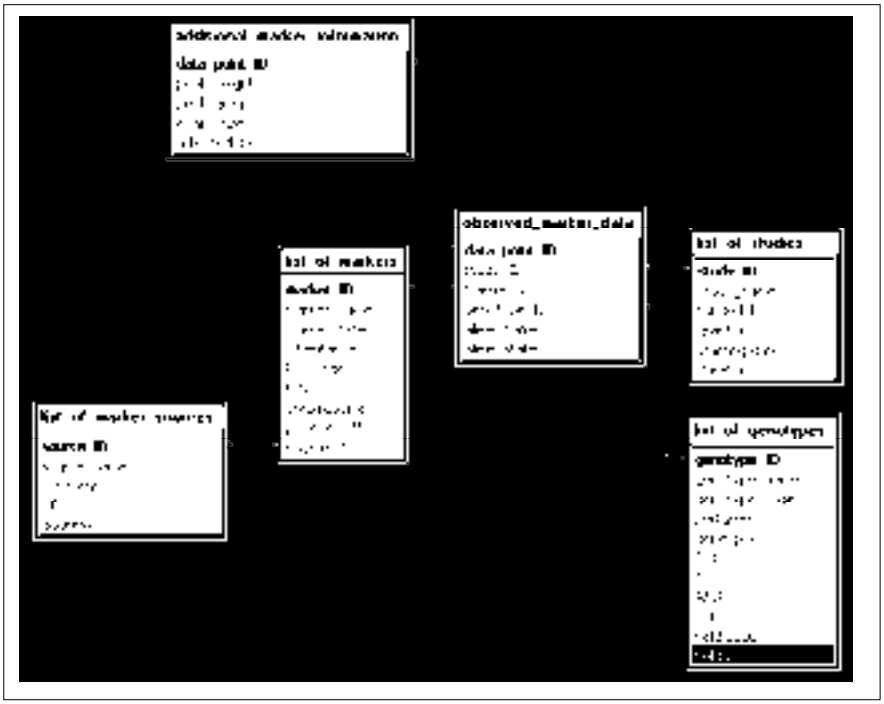

Abbildung 1: **Datenstruktur im Modul "molekulare Daten". Jedes rechteckige Feld steht für eine Tabelle mit Ihren Spalten (Feldern). Linien stehen für Verknüpfungen zwischen Feldern mit identischem Inhalt.**

Tabelle 2: **Struktur der Tabelle "list_of_markers"**

marker_ID	marker_name	marker_type	inheritance	PC_Code*	BIN	chromosome	position	source_ID
…	…	…	…	…	…	…	…	…
1172	bnlg1074	SSR	codominant	*	10.04	10	55,8	0
1173	bnlg1028	SSR	codominant	*	10.05	10	64,6	0
1174	bnlg1839	SSR	codominant	*	10.07	10	83,9	0
1175	bnlg1360	SSR	codominant	*	10.07	10	90,4	0
1176	phi099	SSR	codominant	*	3.02	03	20	2
1177	phi021	SSR	codominant	*	4.03	04	43	2
1178	STAB_A_473	AFLP	dominant	G	1.01	01	1,9	3
1179	STAB_A_475	AFLP	dominant	G	1.01	01	12,6	3
1180	STAB_A_182	AFLP	dominant	I	1.01	01	13,7	3
1181	STAB_A_194	AFLP	dominant	I	1.01	01	17	3
1182	STAB_A_528	AFLP	dominant	H	1.01	01	17	3
1183	STAB_A_195	AFLP	dominant	I	1.01	01	17,1	3
…	…	…	…	…	…	…	…	…

* PC Primer combination

beiden Tabellen auf jeweils die selben Marker bezieht. Dies hat den Vorteil, dass etwa bei einer Änderung des Markernamens oder der Position im Genom nur ein Wert in der Tabelle "list_of_markers" geändert werden muss, nicht aber alle Werte in der Tabelle "observed_marker_data".

In der Tabelle "additional_marker_ information" können zusätzliche Informationen zu den Markerdatenpunkten gespeichert werden, wie die Peakhöhe oder die bei SSRs teilweise bei auf 0,01 bp bestimmte exakte Fragmentgröße. Die Tabelle "list_of_marker_sources" enthält Informationen zu den Firmen, die die Marker zur Verfügung gestellt haben, oder die Datenquellen, die weiterführende Informationen zu den Markern beinhalten, z.B. Links zu Datenbanken im Internet, wie Maize DB (http://www.agron.missouri.edu).

Durch den gezielten Einsatz von Abfragen oder Sichten ist es mit dieser Datenstruktur möglich, alle vorhandenen Daten zu sortieren, zu gruppieren oder gänzlich neue Werte zu berechnen. Auf diese Weise können Inputfiles für die o.a. Anwendungen erstellt oder einfache Analysen innerhalb der Datenbank selbst durchgeführt werden. Beispiele hierfür können unter (http://www.uni-hohenheim.de/~heckenbe/literatur.html) heruntergeladen werden.

2.2 Modul: Felddaten

Zentraler Dreh- und Angelpunkt im Modul Felddaten (*Abbildung 2*) ist die Tabelle "randomization", in der für jede einzelne Parzelle gespeichert ist, welchem Versuch sie zugeordnet ist, welchen Genotyp sie enthält und auf welchem Schlag sie steht. In diese Tabelle können auch zusätzliche Informationen aufgenommen werden, etwa über die Anordnung der Parzellen auf dem Feld oder z.B. bei Pflanzenschutz- oder Düngeversuchen über die individuelle Behandlung der Parzelle.

Die eigentlichen, im Experiment erhobenen Daten werden in der Tabelle "observed_field_data" (*Tabelle 3*) gespeichert, die mit der Tabelle "randomization" über die Parzellenidentifaktionsnummer "plot_ID" verknüpft ist. Hier werden für jede Parzelle alle Daten gespeichert, die routinemäßig in einem Experiment erhoben werden. z.B. Kornertrag, Trockensubstanz, Pflanzenlänge, TKG, Sollpflanzenzahl, etc.

Werden jedoch abweichend von Routineexperimenten viele unterschiedliche und wechselnde Merkmale erhoben, etwa die Bonitur von Registermerkmalen für ein spezielles Experiment oder eine Wertprüfung, empfiehlt es sich, zur Vermeidung von leeren Zellen, diese in einer getrennten Tabelle abzuspeichern, wie in der Tabelle "observed_morpho- logical_ data" beschrieben (*Tabelle 4*). Um Speicher effizienter nutzen zu können und für ein einfacheres Handling der Daten ist nicht mehr jedes Merkmal in einem eigenen Feld gespeichert. Im Feld "trait" ist jetzt das Merkmal, im Feld "parameter_value" der in der Parzelle für das Merkmal erfasste Wert gespeichert, z.B. die Boniturnote oder eine gemessene Größe. Beide Formate lassen sich durch Umgruppieren von Daten problemlos zur gemeinsamen Analyse ineinander überführen. Müssen für ein Merkmal innerhalb einer Parzelle mehrere Datenpunkte erhoben werden, etwa bei Einzelpflanzenbonitur, können diese durch Einfügen einen neuen Feldes z.B. "plant_ID" ein-

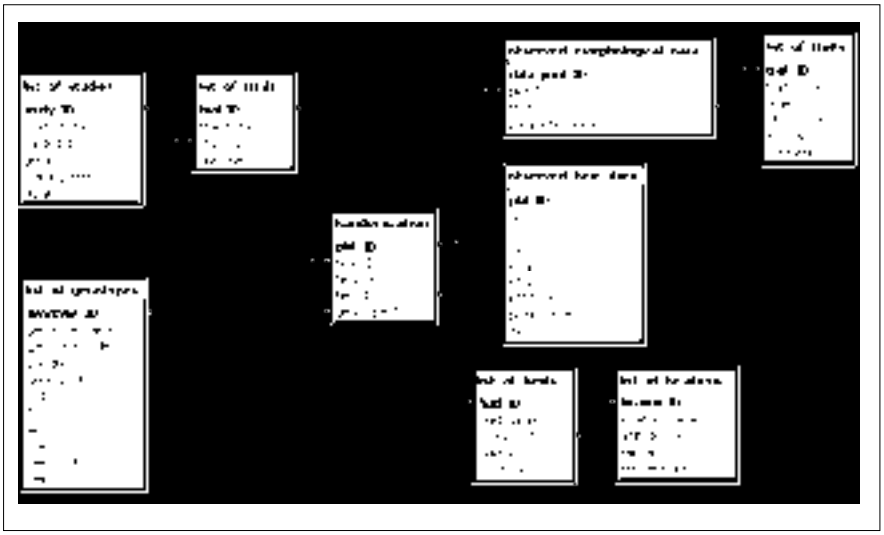

Abbildung 2: **Datenstruktur im Modul "Felddaten"**

Tabelle 3: **Struktur der Tabelle "observed_field_data"**

plot_ID	SP	VP	GP	KFRISCH	KTSE	KTSR	plant_length	TKG
...	...	...	...	...	...	...	...	...
46	26	24	24	2849	242,5	193	225	396
47	26	24	20	2256	246,6	193,6	220	387,2
48	26	26	26	3170	266,6	208,1	125	220
49	26	25	25	1418	249,4	185	230	404,8
50	26	23	19	1296	251,3	191,5	190	334,4
51	26	26	26	1613	255	190,9	170	299,2
52	26	26	26	2437	262,6	206,7	165	290,4
...		...	...	...	...	...	...	...
1168	26	24	24	2061	266,4	190,3	230	404,8
1169	26	22	18	1779	257,9	182	190	334,4
1170	26	22	22	1912	299,8	220,7	170	299,2
1171	26	22	22	3100	294,9	209,2	165	290,4
1172	26	25	21	2753	290,6	205,3	220	387,2
1173	26	25	25	3242	297,9	210,1	230	404,8
1174	26	24	24	1661	299,2	223,3	190	334,4
1175	26	25	21	1450	232	172,7	170	299,2
1176	26	23	23	1499	251,1	187,8	165	290,4
...	...	...	...	...	...	...	...	...

deutig beschrieben werden. Alternativ könnte im Feld "trait" auch nur eine Merkmalsidentifikationsnummer "trait_ID" stehen, und die Informationen zu den Merkmalen selbst in einer eigenen Tabelle "list_of_traits" gespeichert werden.

In den Tabellen "list_of_fields" und "list_of_locations" können Informationen zu den Schlägen und den Versuchsstandorten abgelegt werden, etwa Schlaggröße, Bodenart, Höhe über NN, oder durchschnittlicher Niederschlag. Diese Informationen würden es beispielsweise ermöglichen in einem Experiment über Trockenstress, nur die 3 Schläge mit dem geringsten Niederschlag in die Auswertung mit einfließen zu lassen, oder eine Analyse nur über die wärmsten Versuchsstandorte durchzuführen.

Tabelle 4: **Struktur der Tabelle "observed_morphological_data**

data_point_ID	plot_ID	trait	parameter_value
…	…	…	…
9	2	ASA	5,00
10	2	BTW	77,00
11	2	ANF	3,00
12	2	LOD	1,00
13	2	RLU	23,00
14	2	RLO	18,00
15	2	LSA	13,00
16	2	WUH	170,00
17	2	KAH	60,00
18	2	BBS	7,00
19	2	KLG	117,00
20	5	KDI	35,00
21	5	KRZ	12,75
22	5	KTY	2,00
23	5	FKV	3,00
…	…	…	…

2.3 Modul: Lager und Logistik

Das Modul "Lager und Logistik" (*Abbildung 3*) soll hier am Beispiel der Saatgutverwaltung beschrieben werden. In der Tabelle "list_of_seed_lots" (*Tabelle 5*) sind alle Saatgutpartien mit dem entsprechenden Genotyp, ihrem Gewicht bei Einlagerung (bzw. der Kornzahl) und ihrem Speicherort abgelegt. In der Tabelle "list_of_seed_transfers" (*Tabelle 6*) werden alle Saatgutbewegungen dokumentiert. Über eine Abfrage kann so jederzeit ermittelt werden, wie viel Saatgut von der jeweiligen Partie noch vorhanden ist, ohne bei jedem Saatguttransfer die Daten für das Gewicht oder die Kornzahl der jeweiligen Saatgutpartien ändern zu müssen. Ebenso könnte ein Überwachungsmechanismus eingebaut werden, der etwa bei Unterschreiten einer frei wählbaren Grenze eine Alarmmeldung ausgibt (*Tabelle 7*). Genauso, wie hier am Beispiel der Saatgutverwaltung beschrieben könnte auch eine ähnliche Struktur etwa zur Bevorratung von Chemikalien in Verbindung mit dem Modul "Labor" eingesetzt werden.

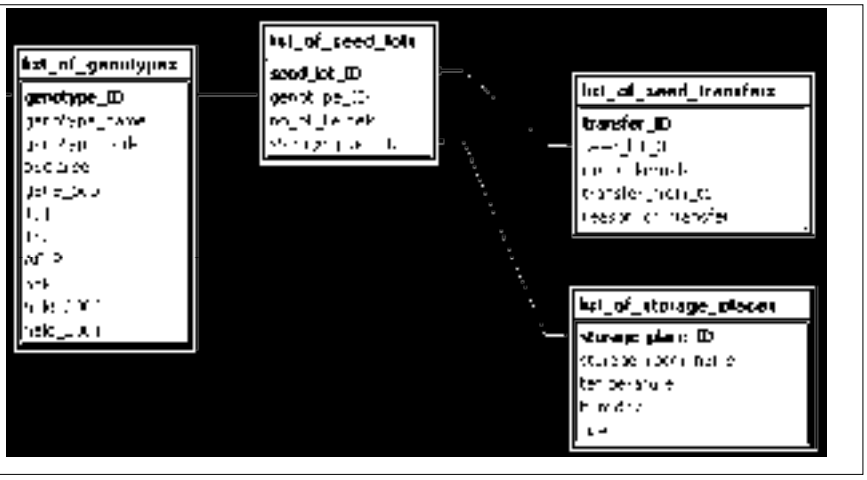

Abbildung 3: **Datenstruktur im Modul "Lager und Logistik"**

2.4 Modul: Labor

Da im Laborbereich die zu verwaltenden Daten stark variieren und daher auch die Struktur der zu verwendenden Datenbanken sehr unterschiedlich ist, müssen Datenbanklösungen für jedes Labor, zumindest aber für jede neu angewandte Technik individuell konzipiert werden. Beispielhaft werden hier in vereinfachter Form die Datenstrukturen zur Durchführung von SSR und AFLP Analysen beschrieben (*Abbildung 4*).

Wie bei den oben beschriebenen Modulen sollten auch im Laborbereich unterschiedliche Daten in getrennten Tabellen gespeichert werden. Es sollte daher eigene Tabellen für PCR-Primer, Restriktionsenzyme, sowie Pufferlösungen oder DNA-Proben geben, die es je nach Fragestellung in geeigneter Weise zu verknüpfen gilt (*Abbildung 4*). Generell zu beachten ist, dass es bei Verknüpfungen mit dem Modul "molekulare Daten" zu keinen Inkonsistenzen, etwa im Feld "marker_ID" kommt.

3. Diskussion

Die hier vorgestellte Datenbankstruktur (*Abbildung 5*) erwies sich bei Ihrer Anwendung als höchst effizient. Die Anwen-

Tabelle 5: **Struktur der Tabelle "list_of_seed_lots"**

seed_lot_ID	genotype_ID	no_of_kernels	storage_place_ID
…	…	…	…
3	2	2000	1
4	145	1000	1
5	146	10000	1
6	123	5000	1
7	1445	7563	1
8	3	1998	1
9	4	2054	1
10	5	2057	1
11	6	5608	1
…	…	…	…

dung des Modells der relationalen Datenbanken und der mittlere Grad der Normalisierung (CODD, 1970) erlaubt die Integration von Daten von kleinen persönlichen Datenbanken in größere Datenbanksysteme, etwa bei der Datenverwaltung von großen Forschungsprojekten mit einer Vielzahl von Teilprojekten und beteiligten Organisationen. Während die Implementierung in größere Datenbanksysteme weitere strukturelle Komponenten erfordern würde (z.B. die konsequente Anwendung von automatisch generierten Primärschlüsseln bei der Verknüpfung von Tabellen oder die Implementierung von Überwachungsmechanismen (="trigger") zur Vermeidung von Dateninkonsistenzen), können die hier vorgestellten Tabellen als Vorlagen für Sichten (="views") zur Datenaufbereitung oder für Eingabeformulare verwendet werden. Dies schafft eine gemeinsame Schnittstelle für Datenaustausch, standardisierte Analysen und kombinierte Analysen von Daten aus verschiedenen Quellen.

Die Flexibilität der hier vorgestellten Datenstruktur und die Möglichkeit einer Integration in größere Projekte ist einer ihrer wesentlichen Vorteile. Im Gegensatz dazu ist die Verwendung von Tabellen aus Tabellenkalkulationsprogrammen oder die Übernahme der Inputformate von Spezialsoftwarepaketen in große Datenbanksysteme schwierig, zeitaufwändig und birgt ein hohes Fehlerrisiko.

Die Möglichkeit, Daten mit anderen, bereits existierenden Daten zu verknüpfen, ermöglicht es dem Versuchsansteller, eine Vielzahl von Analysen durchzuführen, etwa Selektion auf der Basis von mehreren Generationen, oder Diversitätsstudien mit jeweils unterschiedlichen Markersets, z.B. basierend auf Markern unterschiedlicher Chromosomen. Solche Analysen mit getrennt gespeicherten Daten durchzuführen, erfordert einen großen Zeitaufwand zur Dateneditierung und birgt ebenfalls ein hohes Fehlerrisiko (FRISCH et al., 2002).

Das beschriebene Datenbankmodell wurde in abgeänderter und angepasster Form bereits in verschiedenen Studien mit Erfolg zur Datenspeicherung und -aufbereitung eingesetzt:

- einer RFLP-Studie in Mais zum Vergleich der genetischen Diversität von Elternlinien und ihrer abgeleiteten Linien in einem rekurrenten Selektionsprogramm mit Populationen aus Iowa Stiff Stalk Synthetic und Iowa Corn Borer Synthetic No. 1 (HAGDORN et al., 2002);
- einer SSR-Studie zur genetischen Diversität in tropischen Maispopulationen und der Beziehung zwischen genetischer Distanz und Heterosis (WARBURTON et al., 2002; REIF et al., 2002);
- einer Studie zur Untersuchung der Qualität von SSR-Daten von Maisinzuchtlinien aus verschiedenen Stufen der Erhaltungszüchtung, sowie von verschiedenen Züchtern (HECKENBERGER et al., 2002);
- derselben Studie mit AFLP-Daten (HECKENBERGER et al, 2002);
- einer Studie zum Vergleich verschiedener Markersysteme in Weizen (BOHN et al., 1999), und
- mehrere derzeit laufende Studien.

Tabelle 6: **Struktur der Tabelle "list_of_seed_transfers"**

transfer_ID	transfer_date	seed_lot_ID	no_of_kernels	transfer_to_from	reason
…	…	…	…	…	…
3	20.10.2001	3	200	nursery	harvest
4	20.10.2001	4	202	nursery	harvest
5	20.10.2001	5	198	nursery	harvest
6	20.10.2001	6	100	nursery	harvest
7	19.03.2002	3	-20	lab	marker
8	19.03.2002	4	-20	lab	marker
9	19.03.2002	3	-300	nursery	multiplication
10	19.03.2002	4	-300	nursery	multiplication
11	19.03.2002	5	-300	nursery	multiplication
…	…	…	…	…	..

Tabelle 7: **Struktur der Abfrage "current_no_of_kernels"**

seed_lot_ID	no_of_kernels	status
…	…	…
3	1880	ok
4	882	low!
5	9898	ok
6	5100	ok
…	…	…

Die Verwendung der beschriebenen Datenbankstruktur führte darüber hinaus zu beachtlichen Synergieeffekten:

- die Notwendigkeit, der Erstellung von Inputfiles für Spezialsoftwarepakete entfiel;
- die Verwendung der Datenbank in Kombination mit den Standard - Statistiksoftwarepaketen SAS, R, und S-Plus sicherte eine hohe Qualität der biometrischen Auswertungen;
- aufgrund der standardisierten Datenstruktur konnten statistische Analyseroutinen, die für eine Studie programmiert wurden, in anderen Projekten wiederverwendet werden;
- die Datenbank ermöglichte die schnelle Verfügbarkeit von Daten zur Reanalyse und
- die Verwendung von MS Access als Datenbankprogramm ermöglichte allen Mitarbeitern eine schnelle Einarbeitung, dank der ausführlichen Hilfefunktion und den zahlreichen Assistenten, zeigte bei

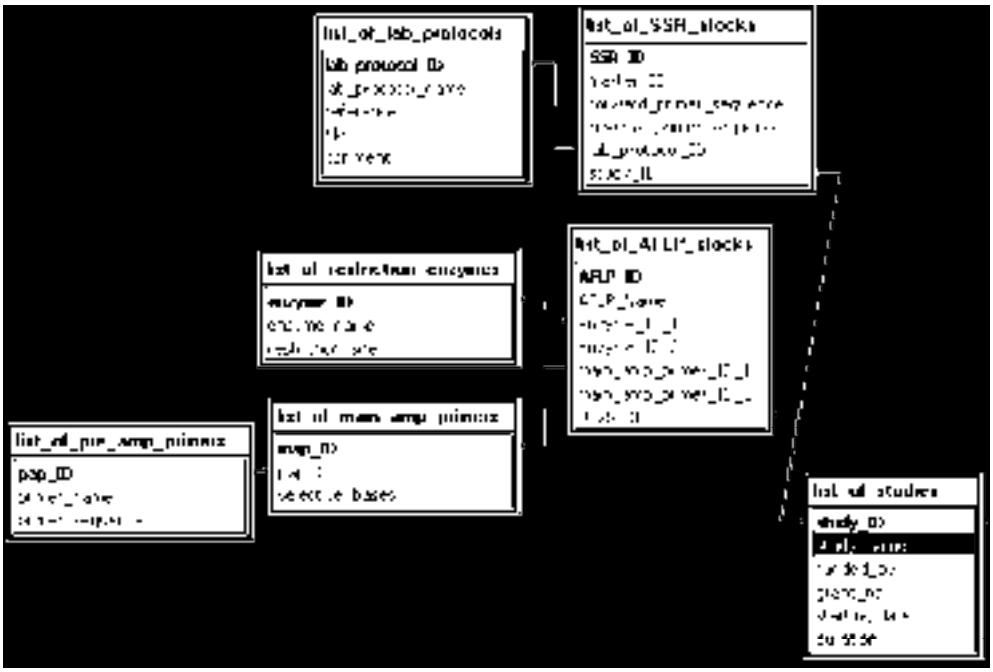

Abbildung 4: **Datenstruktur im Modul "Labor"**

Abbildung 5: **Gesamtstruktur der Datenbank**

großen Datenmengen allerdings Schwächen in der Performance.

Für die in Zukunft weiter steigenden Datenmengen können bei der Verwendung von professionellen Datenbankprogrammen, wie Oracle oder DB2 hochperformante Analysemodule implementiert werden.

Anmerkung: die Verwendung von Firmennamen oder kommerziellen Produkten in diesem Artikel geschieht nur zum Zwecke der Vermittlung wissenschaftlicher Informationen. Dies impliziert keine Empfehlung oder Ablehnung der Produkte oder Firmen.

Literatur

BOHN, M., H.F. UTZ and A.E. MELCHINGER, 1999: Genetic similarities among winter wheat cultivars determined on the basis of rflps, aflps, and ssrs and their use for predicting progeny variance. Crop Sci. 39(1): 228-237.

CHAMBERS, J.M., 1998: Programming with Data. A guide to the S Language. New York: Springer-Verlag.

CODD, E.F., 1970: A relational Model of Data for Large Shared Data Banks. Communications of the ACM 13: 377-387.

FRISCH, M., K.R. LAMKEY and A.E. MELCHINGER, 2002: Storage of molecular marker data in databases for efficient use in plant breeding programs. Zeitschrift fü Agrarinformatik 10: 23-27.

HAGDORN, S., K.R. LAMKEY, M. FRISCH, P.E.O. GUIMARAES and A.E. Melchinger: Molecular Genetic Diversity among Progenitors and Derived Elite Lines of BSSS and BSCB1 Maize Populations. Crop. Sci. in press.

HECKENBERGER, M., M. BOHN, J.S. ZIEGLE, L.K. JOE, J.D. HAUSER, M. HUTTON and A.E. MELCHINGER, 2002: Variation of DNA fingerprints among accessions within maize inbred lines and implications for identification of essentially derived varieties. I. Genetic and technical sources of variation in SSR data. Mol. Breed. 10: 181-191.

HECKENBERGER, M., J. ROUPPE VAN DER VOORT, A.E. MELCHINGER, J. PELEMAN and M. BOHN, 2003: Variation of DNA fingerprints among accessions within maize inbred lines and implications for identification of essentially derived varieties.

II. Genetic and technical sources of variation in AFLP data and comparison with SSR data. Mol. Breed. accepted.

HOLLOWAY, J.L. and S.J. KNAPP: G-MENDEL 3.0 software for the analysis of genetic markers and maps. Oregon State University, Corvallis, Oregon.

IHAKA, R. and R. GENTLEMAN, 1996: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5: 3: 299-314.

LANDER, E.S., P. GREEN, J. ABRAHAMSON, A. BARLOW, M.J. DALY, S.E. LINCOLN and L. NEWBURG, 1987: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genetics 1: 174-181.

REIF, J.C., A.E. MELCHINGER, X.C. XIA, M.L. WARBURTON, D.A. HOISINGTON, S.K. VASAL, M. SRINIVASAN, M. BOHN and M. FRISCH, 2002: Genetic diversity within and between seven tropical maize populations investigated with SSR markers and relation to the heterosis of their crosses. In review.

ROHLF, F. J., 1989: NTSYS-pc Numerical taxonomy and multivariate analysis system. Exeter Publishing Co, Ltd., Setauket, NY.

SAS Institute, 1988: SAS/STAT User's Guide, Release 6.03 edn. SAS.Cary.

SCHNEIDER, S., J.M. KUEFFER, D. ROESSLI and L. EXCOFFIER, 1997: Arlequin: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

UTZ, H.F. and A.E. MELCHINGER, 1996: PLABQTL: A program for composite interval mapping of quantitative trait loci. J. Quant. Trait Loci 2.

WARBURTON, M., X. XIA , J. CROSSA, J. FRANCO, A.E. MELCHINGER, M. FRISCH, M. BOHN and D. HOISINGTON, 2002: Large scale fingerprinting methods for the analysis of genetic diversity of CIMMYT maize germplasm. Crop. Sci.