Abschluß-Bericht

Forschungsverbund: Funktionelle Annotation regulatorischer Genombereiche
Teilprojekt 1: Handhabung und Aufbereitung genregulatorischer Daten

Förderkennzeichen 0311640

Leiter des Teilprojektes

Dr. Edgar Wingender
GBF, AG Bioinformatik
Mascheroder Weg 1
D-38124 Braunschweig

Telefon: 0531 6181-427
Telefax: 0531 6181-266
e-mail: ewi@gbf.de

Beteiligt
I. Liebich, T. Meinhardt, M. Prüß, I. Reuter

Eines der Ziele dieses Bioinformatik-Verbundprojektes war die systematische, funktionelle Klassifizierung regulatorischer Sequenzen (Promotorregionen, Introns, etc.), die Vorhersagen über die physiologische Funktion unbekannter Gene zulassen. Im Teilprojekt 1 "Handhabung und Aufbereitung genregulatorischer Daten" wurden zu diesem Zweck eine Reihe neuer Datenbankressourcen geschaffen.

1. Wissenschaftliche Ergebnisse und andere Ereignisse

Datenbank über Scaffold / Matrix Attached Regions (S/MARt DB)
Ines Liebich

Entsprechend dem Arbeitsprogramm des Braunschweiger Teilprojektes wurde eine Datenbank, die Daten zu S/MARs und daran bindenden Proteinen sammelt und diese insbesondere für biomedizinische und biotechnologische Anwendungen nutzbar macht, aufgebaut. S/MARs (scaffold/matrix attached regions) sind DNA Abschnitte von in der Regel mehreren hundert Basenpaaren Länge, die dauernd oder zeitweilig mit dem unlöslichen Proteinnetzwerk des Zellkerns, der Matrix, verbunden sind.

Das Projekt wurde im Berichtszeitraum (01.01.-31.12.1999) fortgeführt.
Es wurde eine auf Flatfiles basierende Veröffentlichung der gesammelten Daten im Internet angestrebt. Diese Art des Datenveröffentlichung wird u.a. auch von den großen Sequenzdatenbanken SwissProt und EMBL genutzt. Sie entspricht den auf dem HUGO Europe Bioinformatics Forum (1994, Hinxton/Cambridge) gefaßten einschlägigen Empfehlung. Die auf dem Statusseminar im Februar 1999 vorgestellten und diskutierten Entwürfe für solche Flatfiles wurden (von den Teilnehmern) als angemessen betrachtet (siehe Vorjahresbericht).
Zunächst war versucht worden diese Flatfiles mit Hilfe von MS Access zu entwickeln. Damit verband sich die Hoffnung, die gesammelten Daten auf diese Weise schneller zur Verfügung stellen zu können. Leider stellte sich in der weiteren Arbeit heraus, daß mit MS Access erstellte Flatfiles zu intensiv

nacheditiert werden müssen, um den gestellten Anforderungen auch nur annähernd zu entsprechen. Deshalb wurden mehrere Scripte (kleine Computerprogramme) in der Programmiersprache Perl entwickelt, mit denen Flatfiles erzeugt werden können, wie sie im Internet unter der Adresse (http://transfac.gbf.de/SMARtDB/index.html) zugänglich sind. Im Zuge der Veröffentlichung im Internet wurde auch eine Portalseite gestaltet, die es dem Datenbankbenutzer gestattet, durch die Datenbank zu blättern (browsen) oder gezielt nach Einträgen mit bestimmten Inhalt in einzelnen Feldern zu suchen. In der zweiten Hälfte des Berichtszeitraumes wurde die unter MS Access entwickelte relationale Datenbank auf den MS SQL Server portiert. Dieses Datenbankmanagement-system vereinfacht die Abfrage des relationalen Systems durch mehrere Benutzer deutlich. Mit der Portierung wurde S/MARt DB als eigenständiges Modul voll in das TRANSFAC System integriert (Abbildung S. 4). Zur weiteren Pflege der Datenbank wurde ein in Java programmierter und damit prinzipiell plattformunabhängiger Client entwickelt.
Über den gesamten Berichtszeitraum wurde der Datenbestand merklich aufgestockt (Tabelle S. 3). Somit handelt es sich bei S/MARt DB derzeit um die mit Abstand umfangreichste und vollständigste Datenbank zu diesem Gegenstand.

Weitere Entwicklung

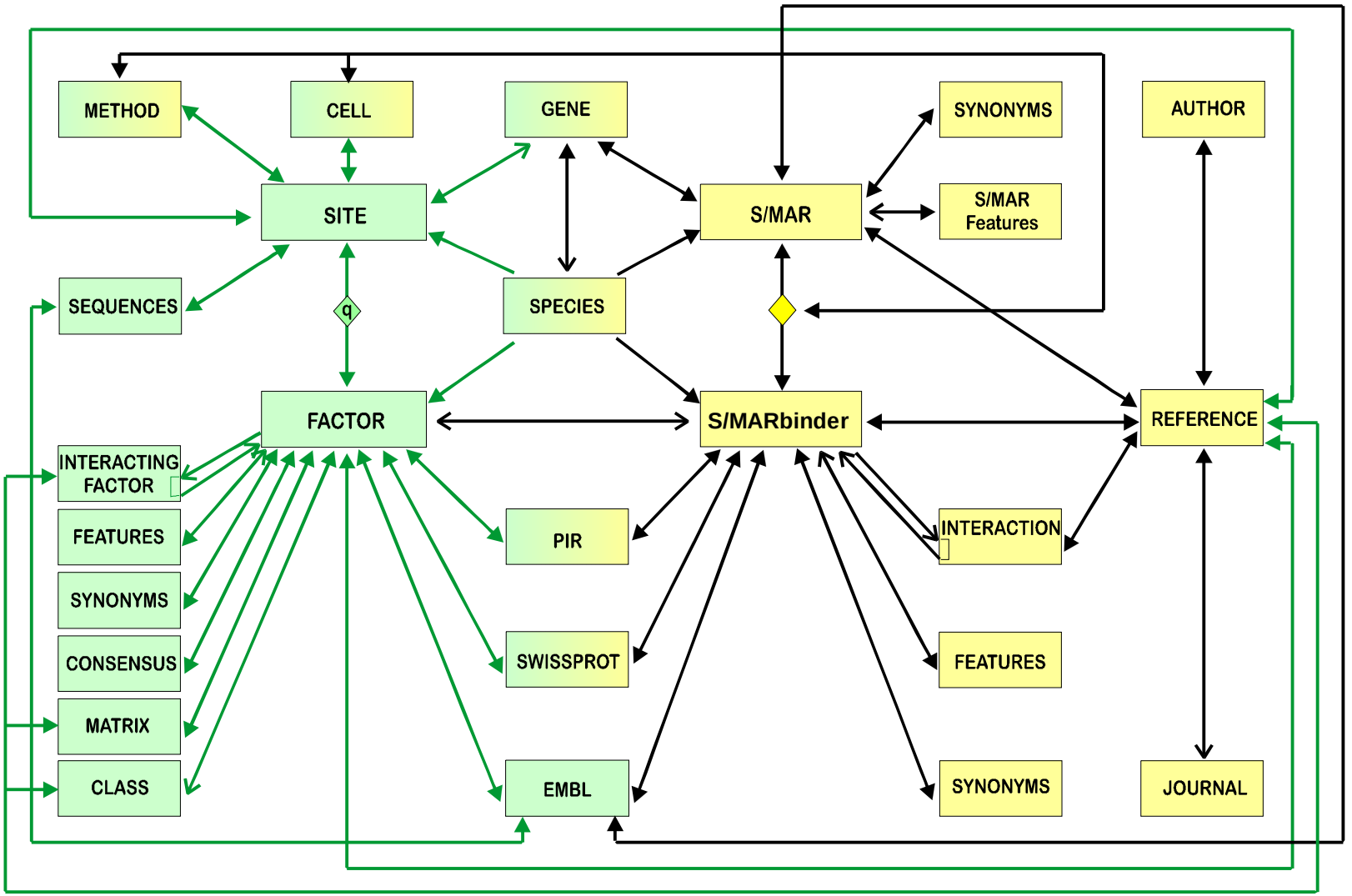
Als nächste Schritte sind folgende Punkte geplant:

- Veröffentlichung der Datenbank mit Hilfe eines Artikels in einer insbesondere von Biologen gelesenen Fachzeitschrift, um S/MARt DB stärker als bisher bekannt zu machen.
- Entwicklung und Erprobung, eines Systems, welches Forschern erlaubt, relevante Daten direkt bei S/MARt DB einzureichen bzw. zu veröffentlichen.
- Umfangreiches und systematisches Data Mining, um typische Sequenzmerkmale von S/MARs zu identifizieren, die dann für die Erkennung derartiger Regionen in genomischen Sequenzen verwendet werden können.

	Datenbestand (intern) zum Zeitpunkt des Statusseminars im Feb. 1999	Datenbestand bei der öffentlichen Vorstellung von S/MARt DB in Internet	Datenbestand 31.12.1999 (intern)
Einträge in der SMAR Tabelle	129	192	229
Einträge in der Gen Tabelle, die mit mindestens einem Eintrag in der SMAR Tabelle verbunden sind	61	110	120
Einträge in der SMARBINDER Tabelle	34	51	57
ausgewertete Veröffentlichungen	85	-	165
in S/MARt DB vertretene Organismen	-	-	28

Tabelle 1: Statistik

Schematic structure of S/MARt DB and its connection to TRANSFAC



Yellow: S/MARt DB tables
Green: TRANSFAC tables

: "one" side of a link
: "many" side of a link

PathoDB – eine modulare Erweiterung von TRANSFAC für die Integration pathologisch relevanter Daten

Thorsten Meinhardt, Manuela Prüß

Die Datenbank PathoDB wurde entwickelt, um Informationen zu pathologischen Formen der eukaryontischen Transkriptionsregulation bereitzustellen. In der Datenbank enthalten sind Informationen zu mutierten Transkriptionsfaktoren und Transkriptionsfaktorbindungsstellen. Dabei umfaßt der Datenbereich sowohl die zugrunde liegende Ebene der molekulargenetischen Information (Genotype; MuSite), die Proteinebene (MuFactor) als auch die Ebene der jeweiligen resultierenden Krankheitsausprägung, die am betroffenen Organismus diagnostiziert werden kann (Phenotype).

Entwicklung der Datenbankstruktur

Zunächst wurde das bereits im Herbst 1998 begonnene Datenbankdesign weitergeführt. Dabei wurde das relationale Datenbankmodell mit der Einführung weiterer Tabellen stark erweitert (Abbildung 1), so daß die Datenbank nun aus insgesamt 34 Tabellen besteht, davon 4 Haupt- und 6 ergänzende Tabellen sowie 24 interne Link-Tabellen. Über 12 weitere Link-Tabellen wurde die Anbindung an TRANSFAC insofern realisiert, als einige auch für den Bereich der mutierten Faktoren und Bindungsstellen wichtigen TRANSFAC-Tabellen – wie Referenzen, Species, Faktorklassen u.a.m. – nun direkt in PathoDB-Tabellen eingelesen werden können. Eine Übersicht über die wichtigsten Tabellen von PathoDB und TRANSFAC sowie über die Tabellen, auf die beide Systeme zugreifen, wird in Abbildung 2 gegeben.

Außerdem wurde eine Reihe von externen Datenbanken über Link-Tabellen an PathoDB angebunden. Dazu gehören EMBL (European Molecular Biological Laboratory Database) und SWISSPROT (Swiss Protein Sequence Database), die Informationen über Gene und Proteine bieten und die an die MuFactor-Tabelle angelinkt sind, OMIM (Online Mendelian Inheritance in Man) und MGI (Mouse Genome Informatics), die Informationen zu ererbten Krankheiten bei Menschen bzw. bei Mäusen bieten und die an die Phenotype- und die Genotype-Tabelle gelinkt sind, sowie HGMD (Human Gene Mutation Database), die an die Genotype-Tabelle gelinkt ist. So wird dem Nutzer der Datenbank der Zugriff auf weiterführende Informationen bestimmter Teilbereiche erleichtert.

Datenlage

PathoDB enthält zur Zeit Informationen zu 10396 mutierten Transkriptionsfaktoren, 19 mutierten Bindungsstellen, 23 Krankheiten und 10415 Genotypen; eine detaillierte Auflistung zeigt Tabelle 1. Die aufgenommenen Transkriptionsfaktoren sind zum Teil solche, die in der pränatalen Entwicklung von Organismen eine wichtige Rolle spielen, wie Pit-1 und Prop-1, die für die Aktivierung von hormonausschüttenden Schilddrüsenzellen zuständig sind, und die verschiedenen Pax-Faktoren, die für die Augen-, z.T. aber auch für die Ohren- und Nierenentwicklung von großer Bedeutung sind. Entsprechend kommt es bei Fehlexpressionen aufgrund von Mutationen zu teilweise schweren Entwicklungsstörungen. Ein weiterer Teil der Faktoren wirkt im mutierten Zustand als Onkogen und verursacht verschiedene Tumorformen, wie dies bei p53 und WT-1 der Fall ist. Die p53-Daten wurden aus einer öffentlich zugänglichen speziellen Datenbank für p53-Mutationen übernommen und im Format an PathoDB angepaßt. Die Eingabemaske für mutierte Faktoren ist in Abbildung 3 dargestellt.

Bei den bisher aufgenommenen mutierten Faktoren und Bindungsstellen handelt es sich hauptsächlich um menschliche, z. T. aber auch um solche aus der Maus. Diese Auswahl bot sich zunächst an, um medizinische und pharmakologische Fragestellungen bearbeiten zu können, muß jedoch keine endgültige sein; prinzipiell könnten alle eukaryontischen Faktoren und Sites berücksichtigt werden.

Abschluß des Teilprojektes „PathoDB“

Die Entwicklung der Datenbank PathoDB wurde mit dem 30.11.99 zunächst abgeschlossen. Seit dem 01.12.99 wird PathoDB von der Firma „BIOBASE – Biologische Datenbanken GmbH“, eine Ausgründung der GBF, weitergepflegt und der Datenbestand ausgebaut, um in Zukunft als Ergänzung zur professionellen Version von TRANSFAC mitvertrieben zu werden.

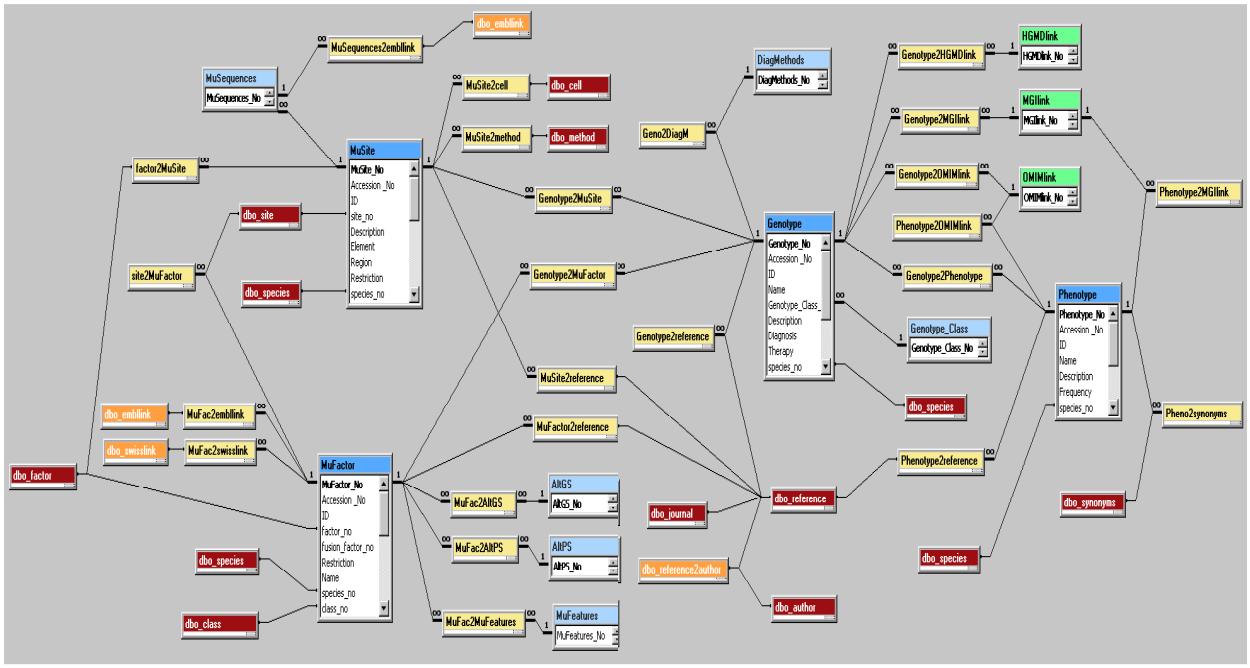


Abbildung 1: Schematische Struktur von PathoDB (ER-Schema). Blau: Haupttabellen; hellblau: Tabellen mit ergänzender Information; rot und orange: Linktabellen zu TRANSFAC; gelb: Linktabellen; grün: Links zu externen Datenbanken.

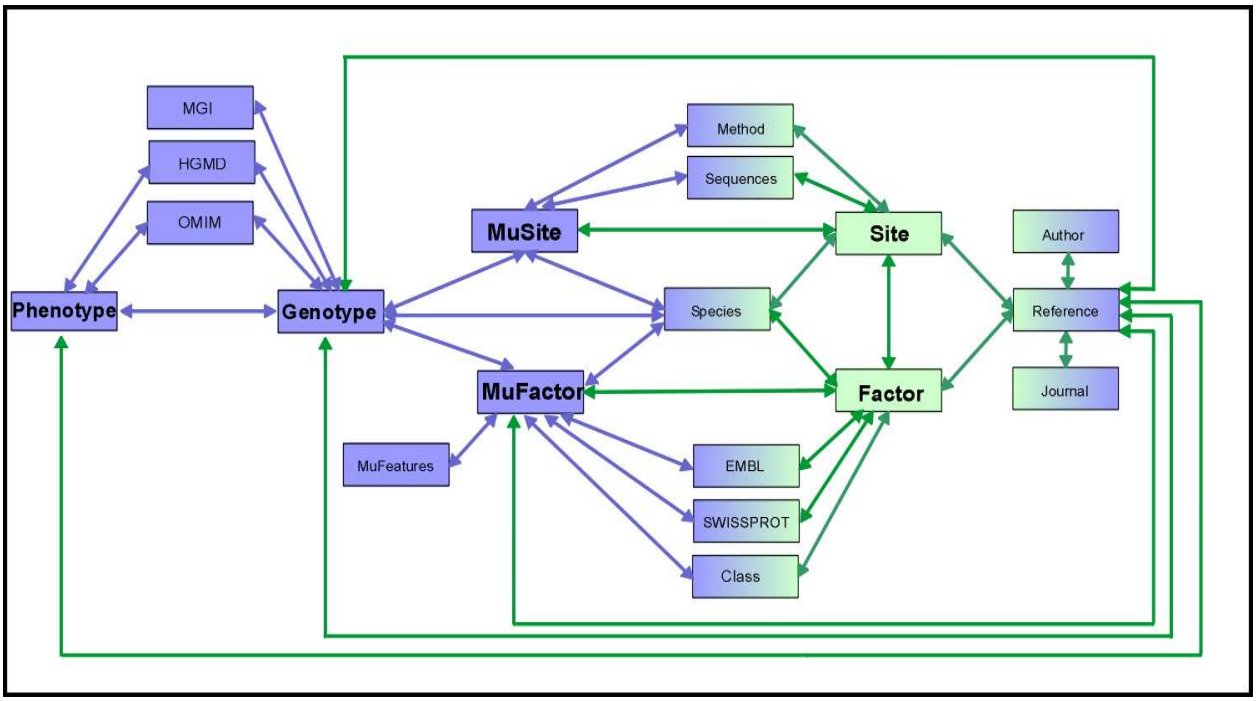


Abbildung 2: Schematische Darstellung einiger ausgewählter Tabellen aus PathoDB und TRANSFAC. Blau: Tabellen in PathoDB; grün: Tabellen in TRANSFAC; blau-grün: Tabellen, auf die von beiden Systemen zugegriffen wird; die Pfeile zeigen die Verbindungen zwischen den Tabellen.

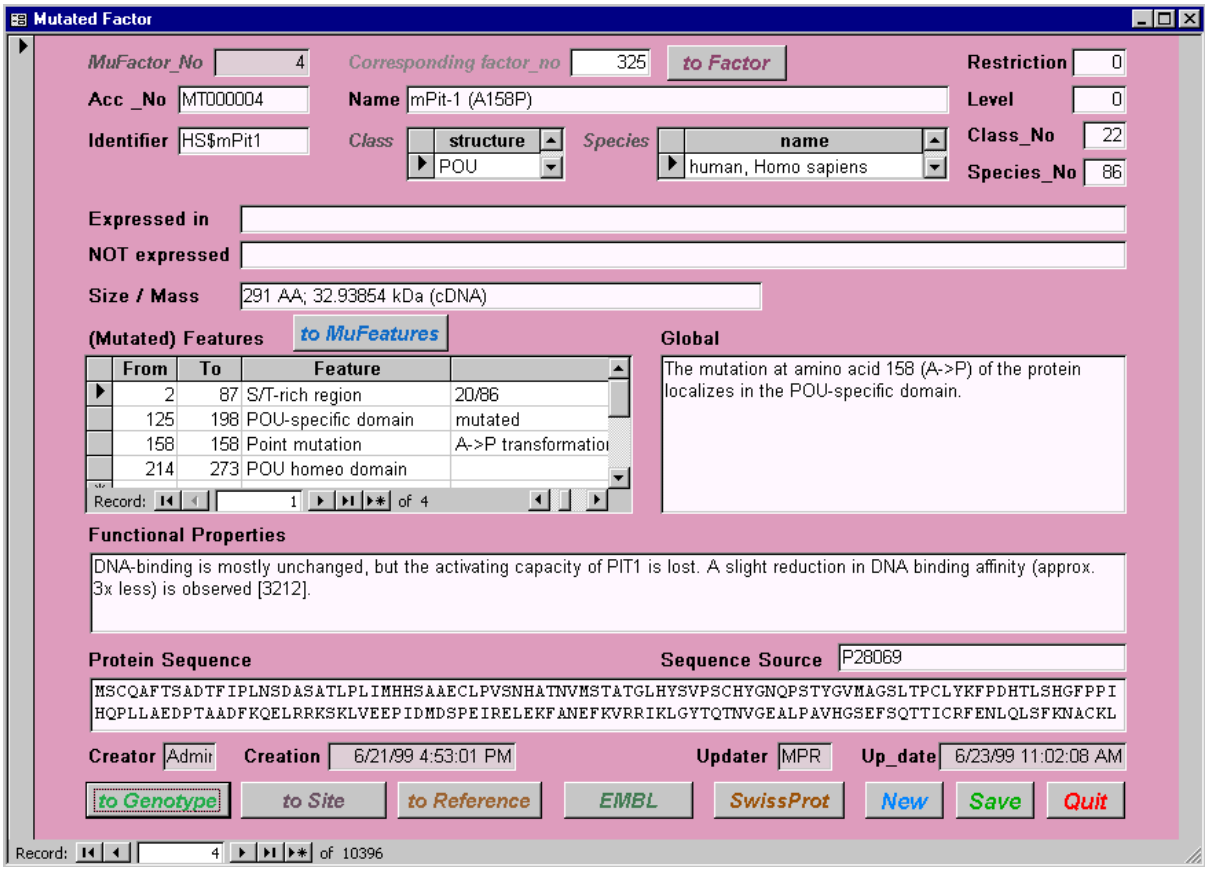


Abbildung 3: Eingabemaske für die PathoDB-Tabelle „MuFactor“ (Bsp.: mPit-1 (A158P))

MuFactors	Anzahl ges.	Maus	Mensch	Phänotyp(en)
Pit-1	9	1	8	Dwarfismus (Maus); Kretinismus (Mensch)
Prop-1	9	1	8	Dwarfismus (Maus); Kretinismus (Mensch)
TTF-2	1	0	1	Congenitaler Hypothyroidismus
Hesx-1	1	0	1	Septo-Optische Dysplasie
Pax-1	2	1	1	“Undulated”-Mutante (Maus); Neuralrohrdefekt (Mensch)
Pax-2	6	0	6	Renal Coloboma-Syndrom
Pax-3	49	2	47	“Splotch-Delayed”-Mutante (Maus); Waardenburg-Syndrom (Mensch)
Pax-6	57	2	55	“Small Eye”-Mutante (Maus); Aniridia, Peters-Anomalie, Foveale Hypoplasie (Mensch)
Pax-8	3	0	3	Congenitaler Hypothyroidismus
p53	10222	0	10222	
WT-1	37	0	37	Wilms-Tumor, Denys-Drash-Syndrom, Mesotheliom, Frasier-Syndrom, Mesangiale Sklerose
MuSites	**Anzahl ges.**	**Maus**	**Mensch**	**Phänotyp(en)**
Beta-Globin	15	0	15	Beta-Thalassämie
Delta-Globin	4	0	4	Delta-Thalassämie

Tabelle 1: In PathoDB enthaltene Datenmengen im Überblick: Mutierte Faktoren (MuFactors) und Bindungsstellen (MuSites), ihre Anzahlen und die resultierenden Phänotypen. (Zu jedem einzelnen mutierten Faktor und jeder Bindungsstelle existiert jeweils ein Genotyp-Eintrag, der hier nicht extra aufgeführt ist.)

In Silico Annotationsdatenbank

Ingmar Reuter

Die In Silico Annotationsdatenbank (ISAD) zur Speicherung und Gliederung von in silico annotierten DNA-Sequenzen basiert in ihrer Struktur auf der TRANSFAC-Datenbank.
Es wurde eine zusätzliche Tabelle für die Verknüpfung der Sequenzen mit den vorhergesagten funktionellen DNA-Elementen aufgenommen (sequence2func_element). Wobei die Tabelle für funktionelle Elemente (func_element) die site-Tabelle aus TRANSFAC ersetzt.
Außerdem.mußte die Verknüpfungstabelle zwischen den funktionellen Elementen (func_element) und der Methoden-Tabelle (method) in erheblichem Umfang erweitert werden, um die Parameter der einzelnen Programme korrekt wiederzugeben.

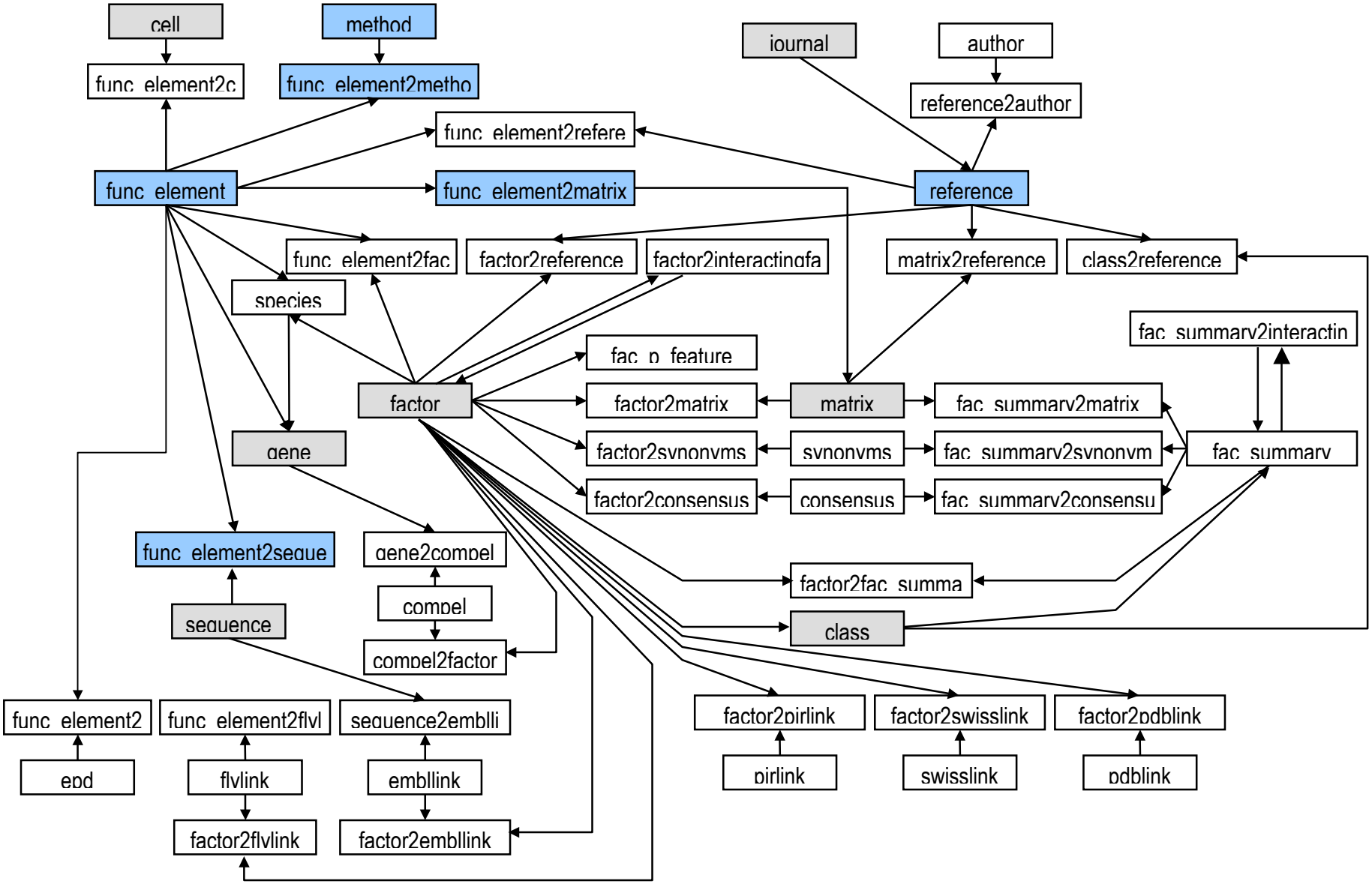
Die Autoren-Tabelle aus TRANSFAC wurde um einige biografische Angaben für den Annotator erweitert.

Die Implementierung erfolgte im Datenbanksystem mysql, das eine einfache Anbindung an ein Web-Frontend ermöglicht. Dieses sollte in PERL implementiert werden.

Die Eingabe der Daten wurde so gestaltet, daß der Output jedes Sequenzanalyse-Programms mit Hilfe eines Output-Parsers in die Datenbank integriert wird. Dieses hat den Vorteil sehr flexibel auf Änderungen reagieren zu können. Eine Alternative wäre die direkte Einbindung der Programme in die Datenbank über das mitgelieferte API. Dazu müssen die Programme allerdings im Quellcode vorliegen und dieser muß entsprechend verändert werden. Dieses ist bei vielen Programmen nicht möglich, da der Quellcode nicht frei verfügbar oder frei änderbar ist. Es ergibt sich durch die direkte Anbindung jedoch ein erheblicher Geschwindigkeitsgewinn, was gerade im Hinblick auf die Analyse sehr großer Datenmengen, wie sie in diesem Projekt avisiert ist, von Vorteil ist.

Die Analyse der Daten erfolgte zunächst noch durch „command line“ gestartete Programme. Dieses soll bei der Umsetzung in die Anwendung noch erweitert werden, sodaß Analysen auch von externen Benutzern durchgeführt werden und in der Datenbank gespeichert werden können.

Die Struktur der Datenbank wird am folgenden vereinfachten ER-Schema deutlich. Die Änderungen gegenüber der TRANSFAC-Datenbank sind blau unterlegt.


cell
method
journal
author
func_element2c
func_element2metho
reference2author
func_element2refere
func_element
func_element2matrix
reference
func_element2fac
factor2reference
factor2interactingfa
matrix2reference
class2reference
species
fac_summary2interactin
fac_p_feature
factor
factor2matrix
matrix
fac_summary2matrix
gene
factor2synonyms
synonyms
fac_summary2synonym
fac_summary
factor2consensus
consensus
fac_summary2consensu
func_element2seque
gene2compel
factor2fac_summa
compel
sequence
class
compel2factor
func_element2
func_element2flvl
sequence2emblli
factor2pirlink
factor2swisslink
factor2pdblink
epd
flvlink
embllink
pirlink
swisslink
pdblink
factor2flvlink
factor2embllink

2. Technologietransfer and die BIOBASE GmbH

PathoDB wurde, wie bereits zuvor schon TRANSFAC, über ein Technologietransferabkommen an die Firma BIOBASE Biological Databases/Biologische Datenbanken GmbH zur wirtschaftlichen Verwertung übergeben. Beide Datenbanken werden dort in vermarktungsfähige Produkte weiterentwickelt und ihr Datenbestand dort weiter gepflegt. Die im Rahmen der öffentlich geförderten Projekte erhobenen Datenbestände bleiben auch weiterhin für Nutzer aus nichtkommerziellen Organisationen frei zugänglich (s. http://www.gene-regulation.de). Die Struktur der In Silico Annotationsdatenbank (ISAD) ist ebenfalls an BIOBASE übergeben worden und wird dort mit Daten gefüllt werden.

Demgegenüber wird de S/MARt Datenbank als Public Domain-Datenbank im Rahmen des Helmholtz-Netzwerks für Bioinformatik (HNB) weitergepflegt, dort wird auch systematisch ein allgemein über das Internet zugänglicher Dateneingabeclient bereitgestellt werden.

Damit wird eine der Vorgaben für das FANGREB-Projekt erfüllt, die hier fortentwickelten zw. neu erstellten Beiträge zu einer internationalen Bioinformatik-Infrastruktur auf eine sich selbst tragende Grundlage zu stellen und damit von einer ständig zu prolongierenden öffentlichen Förderung unabhängig zu machen.

3. Vergleich des Standes des F+E-Vorhabens mit der ursprünglichen Planung

Das Projekt konnte in vollem Umfang erfüllt werden (s. 4.).

4. Erreichung der Ziele

Alle Ziele, die innerhalb des Teilprojektes bearbeitet wurden, wurden voll und ganz erreicht.
Im ursprünglichen Antrag waren die folgenden Ziele formuliert worden:

1. *Für die TRANSFAC-Datenbank wird ein über das WWW bedienbares SQL-Interface entwickelt.*

   Es wurde ein Java-basierter Dateneingabe-Client entwickelt, der derzeit das Standardwerkzeug zur Dateneingabe in die TRANSFAC-Datenbank ist. Er ermöglicht auch das Erstellen von Links zu den großen Sequenzdatenbanken.

2. *TRANSFAC wird um weitere Informationen von regulatorischer Bedeutung erweitert.*

   Es wurde ein vollständig neues TRANSFAC-Modul geschaffen, das Informationen über scaffold/ matrix attached regions (S/MARs) wie deren Lokalisierung, Sequenzen, und funktionell interagierende Proteine bereitstellt. Dieses Modul, die S/MARt DB (*S/MAR transaction database*) wurde über das Internet öffentlich verfügbar gemacht (http://transfac.gbf.de/SMARtDB/ index.html sowie unter http://www.gene-regulation.de/).

3. *Die Datenbank wird gezielt um solche Informationen erweitert, die von medizinischem Interesse sind (pathologische Aberrationen in cis-Elementen und in den Transkriptionsfaktoren).*
   Hierzu wurde eine eigenständige neue Datenbank geschaffen (PathoDB), die zugleich mit TRANSFAC voll integriert wurde. Der im Projekt akquirierte bzw. aus der öffentlichen p53-

Datenbank importierte Teil wurde über das Internet öffentlich zugänglich gemacht (http://www.gene-regulation.de/).

4. *Verschiedene verfügbare Sequenzanalyseprogramme, z. B. das in der Gruppe entwickelte System zur Clusteranalyse, aber auch Entwicklungen anderer Gruppen innerhalb und außerhalb des Verbundes werden mit der Datenbank verknüpft.*

   Nach einer umfassenden Bestandsanalyse wurde das Schwergewicht auf Matrix-basierte Sequenzanalyseprogramme gelegt. Unter Zugrundelegung des bereits früher gemeinsam mit der GSF entwickelten Programms MatInspector wurden systematisch positive und negative Trainingssätze zusammengestellt und zur individuellen Optimierung der Parametrisierung dieser Analyseroutine für sämtliche in der TRANSFAC-Datenbank verfügbaren Suchmuster verwendet.

5. *Es wird eine TRANSFAC-analoge Datenbank mit in silico-annotierten genomischen Regulationssignalen aufgebaut und öffentlich verfügbar gemacht.*

   Es wurde eine ISADB (In Silico Annotation Database) aufgebaut, die zur Vorhaltung kontextualer Eigenschaften regulatorischer Elemente geeignet ist.

5. Notwendige Veränderungen der Zielsetzungen

Keine

6. Angemeldete Schutzrechte

Keine, da Datenbanken durch Copyright sowie ein Schutzrecht *sui generis* geschützt sind.

Publikationsverzeichnis

Ergebnisse aus dem FANGREB-Projekt sind in folgende Publikationen eingegangen:

1. Heinemeyer, T., Chen, X., Karas, H., Kel, A. E., Kel, O. V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F. and Wingender, E.: Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. Nucleic Acids Res. 27, 318-322 (1999).
2. Kel, A., Kel-Margoulis, O., Babenko, V. and Wingender, E.: Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. J. Mol. Biol. 288, 353-376 (1999).
3. Chen, X., Dress, A., Karas, H., Reuter, I. and Wingender, E.: A database framework for mapping expression patterns. Computer Science and Biology. Proceedings of the German Conference on Bioinformatics (GCB '99). R. Giegerich, R. Hofestädt, T. Lengauer, W. Mewes, D. Schomburg, M. Vingron and E. Wingender (eds.). GBF-Braunschweig and University of Bielefeld, pp. 174-178 (1999).
4. Kel-Margoulis, O. V., Romashchenko, A. G., Kolchanov, N. A., Wingender, E. and Kel, A. E.: COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. Nucleic Acids Res 28, 311-315 (2000).
5. Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüß, M., Reuter, I. and Schacherer, F.: TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. 28, 316-319 (2000).