Vergleichende Analyse und Erkennung genregulatorischer Nukleinsäuresequenzen (GENUS)

Im Rahmen des BMBF-Förderprogramms
"Molekulare Bioinformatik"
gefördertes Verbundprojekt
1993-1996

- Abschlussbericht -

Beteiligte:

Edgar Wingender, *Abt. Genomanalyse, Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-38124 Braunschweig (Koordination)*
Teilprojekt: Entwicklung der Datenbank TRANSFAC zur Basis für umfassende Analysen regulatorischer Nukelinsäure-Sequenzen

Andreas Dress, *Fakultät für Mathematik, Universität Bielefeld, Universitätsstr., D-33615 Bielefeld*
Teilprojekt: Clusteranalytische Verfahren zur Klassifikation und Identifikation genregulatorischer Bereiche

Heinz Sklenar, *Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Str. 10, D-13125 Berlin-Buch*
Teilprojekt: Berechnung und Analyse sequenzabhängiger Raumstrukturen genregulatorischer DNA-Fragmente

Thomas Werner, *AG BIODV, GSF - Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstr. 1, D-85758 Oberschleißheim*
Teilprojekt: Entwicklung von Verfahren zur vollständigen Beschreibung und Abgrenzung regulatorischer Nukleinsäure-Sequenzen

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung und Technologie unter dem Förderkennzeichen 01 IB 306 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Inhalt

Vorbemerkung: *Die schwerpunktmäßigen Beiträge der beteiligten Arbeitsgruppen sind folgendermaßen gekennzeichnet: [1], GBF; [2], GSF; [3], MDC; [4], Universität Bielefeld.*

Kurzfassung

Im Rahmen des GENUS-Verbundprojektes wurde eine Reihe von Werkzeugen für die Charakterisierung und Identifizierung genregulatorischer Genomelemente entwickelt, integriert und der wissenschaftlichen Öffentlichkeit zugänglich gemacht. Zu dem Paket dieser Entwicklungen gehören eine Reihe von Bausteinen, die weltweit einzigartig sind und international ein dementsprechendes Interesse gefunden haben. Insbesondere sind hier die Datenbank TRANSFAC (GBF), ein Bündel an Sequenzanalyseprogrammen zur Einzelement- wie zur Kontextanalyse (GSF), eine Bibliothek von DNA-Strukturparametern (MDC) sowie neue Algorithmen zur Verwandtschafts- und Korrelationsanalyse (Universität Bielefeld) zu nennen.

1. Einleitung

Die Regulation der Genexpression findet im wesentlichen, nach heutigem Kenntnisstand sogar zum größten Teil auf Transkriptionsebene statt. Dazu besitzt jedes eukaryote Gen außer seiner kodierenden Region noch regulatorische Bereiche (Promotor, Enhancer u. a. m.). Diese sind modular aus einzelnen Elementen von ca. 5-25 Basenpaaren zusammengesetzt, an die *trans*-aktivierende (oder -reprimierende) Proteine, die sogenannten Transkriptionsfaktoren, binden. Von diesen Konstituenten genregulatorischer Prozesse ist inzwischen ein beträchtliches Repertoir bekannt. Durch die additive, synergistische oder antagonistische Wirkung der einzelnen DNA-Elemente, durch eine Vielzahl möglicher Interaktionen der Transkriptionsfaktoren miteinander und durch ihre posttranslationale Modifizierung ergibt sich eine außerordentliche kombinatorische Vielfalt, die das gesamte Spektrum benötigter Regulationsmechanismen abdeckt (Wingender, 1993).

Im Rahmen der anlaufenden Genomforschungsprojekte, insbesondere der internationalen Bemühungen um die Entschlüsselung des menschlichen Genoms, ist mit einer enormen Menge an rohen Sequenzdaten zu rechnen. So umfaßt etwa das menschliche Genom 3×10^9 Basenpaare (bp). Um aus diesen Daten nutzbare genomische Information zu gewinnen, bedarf es der funktionellen Annotation. Diese muß sich auf die transkribierten Bereiche, d. h. die offenen Leseraster und somit die Intron/Exon-Übergänge, auf die Genprodukte und ihre funktionelle Charakterisierung, und nicht zuletzt auf die Regulation der Gene beziehen. Nur

über die Kenntnis seiner Regulation läßt sich ein Gen in das komplexe Netz biologischer Funktionsbeziehungen einordnen.

Es stellt sich somit das grundlegende Problem, einzelne regulatorische Sequenzelemente so zu charakterisieren, daß sie auch in neu ermittelten genomischen Sequenzen eindeutig identifiziert werden können. Darauf sind dann weiterführende Methoden zur Erkennung komplexerer Strukturen wie zusammengesetzter Elemente (composite elements), Promotoren/Enhancer sowie ganzer regulatorischer Domänen im Chromatin aufzusetzen.

Für den Einsatz in Genomprojekten ist die Fähigkeit zur Handhabung sehr langer Sequenzen unabdingbar. Die im Verbundprojekt entwickelten Tools wurden daher jeweils an den längsten bekannten Sequenzen, dem komplett sequenzierten Hefegenom, untersucht.

2. Die Datenbasis

2.1 Struktur [1]

Bevor mit einer systematischen und umfassenden Analyse regulatorischer Elemente begonnen werden kann, bedarf es einer Sichtung und Sammlung der entsprechenden Daten aus der Literatur. Zu diesem Zweck wurde eine Datenbank (TRANSFAC) aufgebaut, die die relevanten Informationen über genomische Regulationssignale (relativ kurze Nukleotid-Sequenzen) und die an sie bindenden Proteine enthält. Sie konnte während der Laufzeit des Verbundprojektes qualitativ wie quantitativ deutlich erweitert werden. Sie wird als relationales Datenbanksystem vorgehalten und gepflegt (Abb. 1; Knüppel et al., 1994; Wingender, 1994). Ihre zentrale "Achse" ist die n:m-Beziehung zwischen den Tabellen SITE, die hauptsächlich die genomischen Positionen von Regulationssequenzen enthält, und FACTOR. Die FACTOR-Tabelle enthält Informationen über die physiko-chemischen und biologischen Eigenschaften der Transkriptionsfaktoren (Länge, kalkulierte und experimentell ermittelte Molekulargewichte, lokale und globale strukturelle Charakteristika, funktionelle Eigenschaften, Expressionsmuster); als letzte Erweiterung wurden die vollständigen Proteinsequenzen der Transkriptionsfaktoren mit aufgenommen. Dieser Schritt hatte sich als notwendig erwiesen, da in den Proteinsequenz-Datenbanken häufig die zahlreichen Spleißvarianten nicht aufgenommen sind oder nur indirekt in den Eigenschaftstabellen beschrieben werden, oder sie fehlen vollständig, weil die entsprechenden offenen Leseraster in den Nukleinsäure-Datenbanken nicht annotiert sind.

Darüber hinaus geben weitere Tabellen zusätzliche Informationen, die den Nutzer in die Lage versetzen, die experimentelle Evidenz und die Bedeutung der abgerufenen Information einzuschätzen; dazu gehören Angaben über die zelluläre Umgebung, in der eine bestimmte regulatorische DNA-Protein-Wechselwirkung nachgewiesen wurde (CELL-Tabelle), die experimentellen Methoden dieses Nachweises (METHOD), die biologische Spezies, von der das untersuchte Gen stammt (SPECIES), und die eigentliche Sequenzinformation (SEQUENCE), die alle mit der SITE-Tabelle verknüpft sind (s. Abb. 1). Die Tabelle GENE gibt übergeordnete Informationen über die Gene, denen einzelne Regulatorelemente zugeordnet sind.

Eine ähnliche, übergeordnete Funktion nimmt für die FACTOR-Tabelle die Tabelle CLASS ein, die die strukturellen Prinizipien der DNA-Bindungsdomänen der Transkriptionsfaktoren beschreibt. FACTOR ist zudem mit einer Liste von SYNONYMS und einer über interagierende Faktoren verbunden. Wichtig für die Sequenzanalyse sind die Tabellen MATRIX und CONS. MATRIX enthält Nukleotidverteilungsmatrizen, die entweder aus random selection-Untersuchungen oder aus publizierten bzw. selbst vorgenommenen Zusammenstellungen von Bindungsstellen einzelner Faktoren abgeleitet wurden. Sie werden vom Programm MatInspector (s. u.) für die Sequenzanalyse verwendet. CONS enthält alle relevanten Angaben von Consensus-Beschreibungen, wie sie vom Programm ConsInd (Frech et al., 1993) für kompilierte Bindungsstellen errechnet wurden (s. u.). Ferner gibt es Verknüpfungen mit externen Datenbanken (EMBL, SwissProt, PIR, Flybase, PROSITE, EPD; s. u., Abschnitt 2.3).

Während der Projektlaufzeit ist der Zugriff über das WWW zum unbedingten “State of the Art” geworden. Zuvor war in regelmäßigen Abständen eine reine ASCII flat file Version exportiert und verteilt worden. Das wird nach wie vor durchgeführt, die flat files bestehen inzwischen aus 6 Tabellen (SITE, FACTOR, CELL, CLASS, MATRIX, GENE), die Information der übrigen Tabellen wird in diese integriert, so daß sich die Struktur vereinfacht (Abb. 1b). Diese flat files werden in html-Format umgewandelt und über geeignete Nutzeroberflächen über das WWW zur Verfügung gestellt (Wingender et al., 1996, 1997).

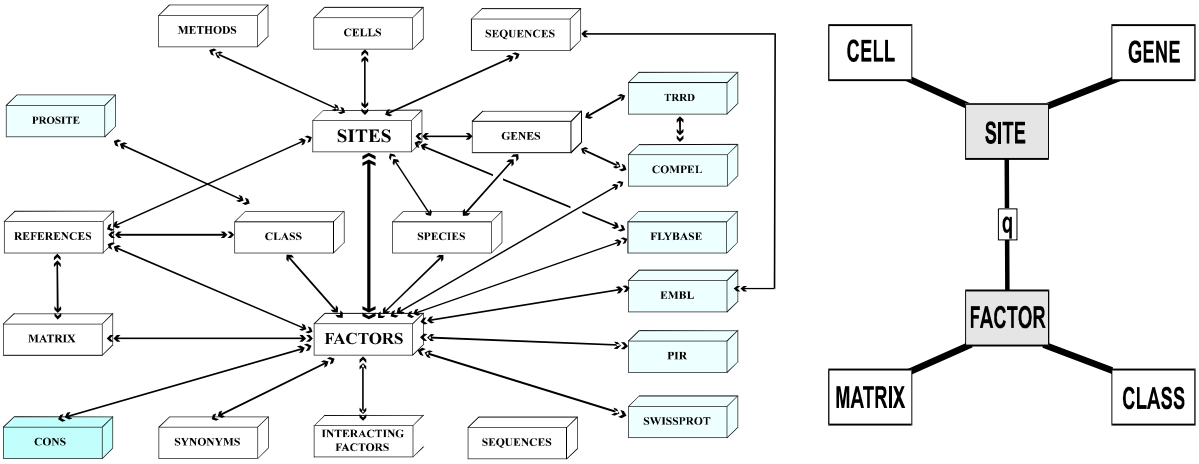


Abb. 1: Relationales Modell der TRANSFAC Datenbank (rechts) und vereinfachte Struktur der WWW-Version (links). Die Namen der einzelnen Tabellen sind in den Kästen angegeben, Pfeile mit zwei Doppelköpfen repräsentieren n:m-Beziehungen im relationalen Modell, solche mit einem Einfach- und einem Doppelkopf 1:n-Beziehungen. Die grau schattierten Tabellen deuten externe Datenquellen an. Die GENE-Tabelle ist identisch mit GENE in der Datenbank TRRD (s. Text).

2.2 Umfang [1]

Der noch während der Projektlaufzeit vorbereitete TRANSFAC release 3.1 enthält z. Z. Informationen über mehr als 4000 genomische Sites, die über 800 Genen zugeordnet sind (Tab. 1). In der SITES Tabelle finden sich zudem über 300 synthetische Sequenzen sowie über 200 Consensus-Sequenzen im 15-Buchstaben-IUPAC-Code. Gegenüber der ersten Auflistung regulatorischer Elemente ist der Umfang der Datenbasis um etwa eine Größenordnung angewachsen, viele Eigenschaften sind völlig neu hinzugekommen.

In den letzten TRANSFAC-Releases ist insbesondere die FACTOR-Tabelle stark angewachsen (um ca. 33% bei der Zahl der Einträge, um ca. 40% bei der Gesamtmenge der Information in release 3.1 gegenüber 2.4). Hier wird langfristig auch ein Schwerpunkt bei der Datenbankpflege liegen, da die Informationen dieser Tabelle regulatorische Mechanismen beschreiben.

Tab. 1: Inhalt und Umfang der Datenbank TRANSFAC.

	r3.1 (02/97)	compil. ('88)[1]
SITES	**4362**	464
Sequences	4105	395
Genes	1041	122
EMBL x-ref	4055	-
synthetic sequences	316	-
consensus sequences	238	-
FACTORS	**2080**	145
Class assignments	1108	15
EMBL x-ref	1933	-
SwissProt	888	-
PIR	955	-
CELLS	**839**	101
CLASS	**28**	2
MATRIX	**258**	-
METHODS	**52**	15
REFERENCES	**5123**	209
Jounals	115	12

[1] Compilation von Wingender (1988).

2.3 Verknüpfungen mit externen Datenquellen [1]

Von zentraler Bedeutung ist die Entwicklung eines föderierten Datenbanksystems mit TRRD (Transcription Regulatory Region Database) und COMPEL vom Institute of Cytology and Genetics (ICG) in Novosibirsk. TRRD enthält Informationen über komplette regulatorische Regionen und ist daher komplementär zur TRANSFAC Datenbank, deren Focus auf den Einzelelementen und den an sie bindenden Faktoren liegt (Kel et al., 1996; Wingender et al., 1997). COMPEL ist eine Unterdatenbank von TRRD (Transcription Regulatory Region Database; ICG, Novosibirsk, Rußland), welche die Zusammensetzung und Funktionalität von sogenannten "composite elements" beschreibt, also solchen Sequenzen, die durch die Kombination mehrerer einzelner Bindungsstellen eine neue Einheit mit einer neuen regulatorischen Qualität erzeugen (Kel et al., 1995). Für die Föderation dieser Datenbanken stellt die bereits weiter oben erwähnte Implementierung einer gemeinsamen GENE-Tabelle den ersten Schritt dar, dem eine weitere intensive Vernetzung durch zahlreiche Querbezüge zwischen den verschiedenen Tabellen aller drei Datenbanken folgen werden. Für die gemeinsame Pflege der GENE-Tabelle wurden mit den russischen Partnern detaillierte Absprachen, z. B. über die Vergabe der Accession Numbers, getroffen.

Darüber hinaus sind die meisten TRANSFAC SITE- und FACTOR-Einträge mit der EMBL-Datenbank sowie die *Drosophila*-Einträge mit FlyBase verknüpft, FACTOR darüber hinaus mit SwissProt und PIR, CLASS mit PROSITE (s. a. Abb. 1). In den flat files und somit am Web finden sich die Datenbank-Querverweise mit Accession Number und Identifier (bzw. Eintrag-Namen oder Äquivalent), in den html files in Form aktiver Hyperlinks. Darüber hinaus enthalten die SITE - EMBL Verknüpfungen die Positionsangaben für das regulatorische Element in der entsprechenden EMBL-Sequenz, sowie in den FACTOR-EMBL-Verknüpfungen die Angabe, ob es sich um eine genomische oder eine cDNA-Sequenz für den zugeordneten Transkriptionsfaktor handelt. Umgekehrt wurden Referenzen auf TRANSFAC in der EMBL-Datenbank, in SwissProt und in Flybase aufgenommen.

Jeder neue Datenexport zur Erstellung von einem neuen ASCII flat file release wird einer Konsistenzprüfung mit EMBL und SwissProt unterzogen. Die Verknüpfungen mit FlyBase werden von dort vorgenommen.

2.4 Verfügbarkeit [1]

Die relationale TRANSFAC-Version muß lokal installiert werden, per anonymous ftp können verschiedene Datenbankmanagement-Systeme vom Server der GBF abgeholt werden (ftp.gbf-braunschweig.de). Als ASCII flat files wird TRANSFAC mit der CD ROM des EBI zusammen mit der EMBL-Datenbank, SwissProt und anderen verteilt. Auch in dieser Form kann es vom GBF Server oder dem des EBI bezogen werden.

Im Laufe der letzten Jahre hat sich das WWW als nutzerfreundlicher Standard auch für Datenbankzugriffe entwickelt. TRANSFAC wurde daher 1995 auch auf diesem Weg zugänglich gemacht. Zur Zeit sind bis zu 15.000 Zugriffe pro Woche zu verzeichnen (http://transfac.gbf-braunschweig.de). Hierbei sind die Zugriffe auf verschiedenen Mirror Sites nicht mitgerechnet, die inzwischen in Peking, Novosibirsk und St. Petersburg eingerichtet wurden, oder andere TRANSFAC-Kopien etwa an der Pennsylvania University, wo eigene Suchroutinen entwickelt wurden.

Darüber hinaus wurde TRANSFAC in datenbankübergreifende Retrieval-Systeme eingebunden. SITES und FACTORS stehen als eigenständige Datenbanken im SRS (Sequence Retrieval System; T. Etzold, EMBL Heidelberg) zur Verfügung, TRANSFAC als ganzes im DBGET (Y. Akiyama, University of Kyoto) (Abb. 2a). Spezielle Zugriffsmöglichkeiten wurden durch Verbindung mit den Patternsuchroutinen des GCG program package am EMBL (Heidelberg) und am Weizmann Institute of Science (Rehovot, Israel) geschaffen (Abb. 2b).

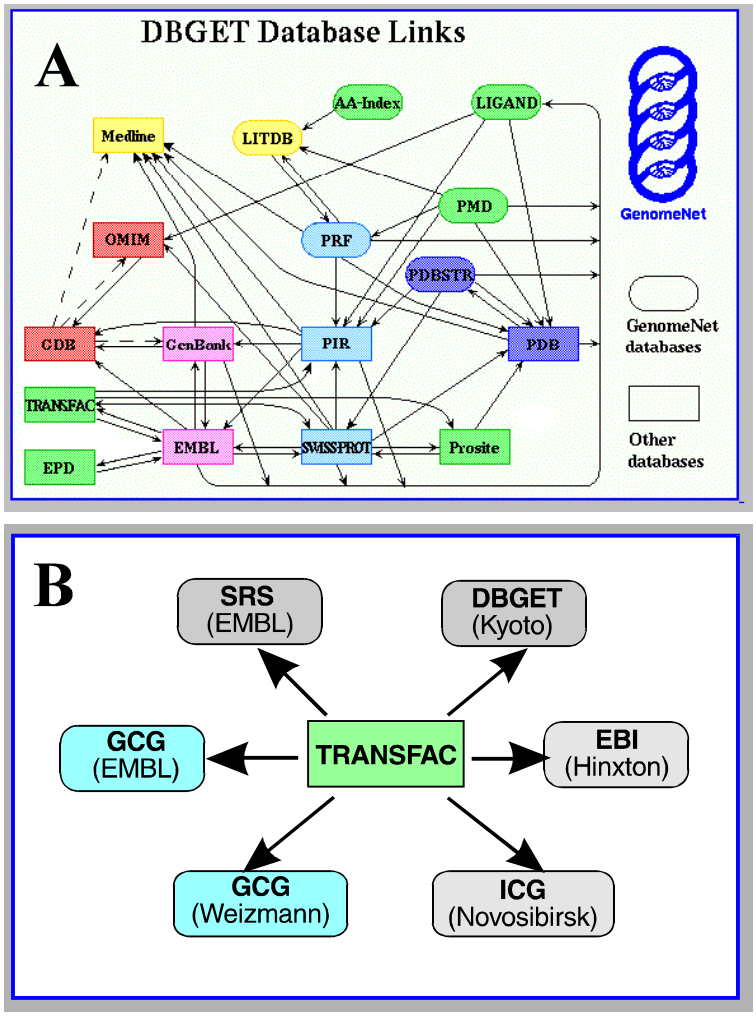


Abb. 2: (a) Einbindung der TRANSFAC Datenbank in das datenbankübergreifende Zugriffssystem DBGET (Kyoto, Japan). U. a. sind die Vernetzungen von TRANSFAC mit der EMBL Datenbank und SwissProt (wechselseitig) sowie mit PIR und PROSITE (einseitig) zu erkennen. (b) Bezug auf / Nutzung von TRANSFAC-Daten durch andere Datenquellen und Nutzersysteme. Neben der Einbindung in SRS und DBGET wird TRANSFAC durch verschiedene Resourcen am EBI (Hinxton) und am ICG (Novosibirsk) genutzt und wurde durch Zusatzentwicklungen am EMBL (Heidelberg) und am WIS (Rehovot) an das GCG Program Package angebunden.

3. Tools zur statistischen Beschreibung regulatorischer Elemente

3.1 PatternSearch [1]

Die in der SITE-Tabelle enthaltenen Sequenzinformationen regulatorischer Genomelemente können zur Analyse neu ermittelter genomischer Sequenzen auf potentielle Transkriptionsfaktor-Bindungsstellen herangezogen werden. Dazu wurde ein Programm entwickelt, das wie die Datenbank über ein WWW-Interface zugänglich ist (http://transfac.gbf-braunschweig.de/patSearch/patsearch.pl). Es beruht auf einer binären Kodierung der Test- wie der Datenbanksequenzen, was einen sehr schnellen Vergleich insbesondere unter Einbeziehung degenerierter Nukleotid-Codes ermöglicht. Es macht zudem Gebrauch von der Notation vieler TRANSFAC-Sequenzen, die für besonders wichtig erkannte Core-Positionen von weniger wichtigen Randpositionen durch Groß- und Kleinschreibung unterscheidet. Dementsprechend kann der Nutzer die Zahl der zulässigen Mismatches innerhalb und außerhalb der Core-Region einstellen (Wingender et al., 1996b). Darin unterscheidet sich PatternSearch von der im MatInspector eingebundenen allgemein gehaltenen IUPAC Suchroutine.

3.2 MatInd/MatInspector [2, 1]

Um auf Nukleotidverteilungsmatrizen, die sich aus der Vielzahl der in TRANSFAC vorhandenen Bindungsstellen einzelner Faktoren ergeben, schneller zugreifen zu können, aber auch zur Nutzung von Matrizen, die aus random selection-Untersuchungen gewonnen wurden und für die es naturgemäß keine Sequenzumgebungen gibt, wurden von ConsInd/ConsInspector (s. u., 3.3) Programme abgeleitet, die mit eng begrenzten Matrizen arbeiten. Da hier nur die eigentlichen Erkennungssequenzen betrachtet werden, entfällt beim MatInd gegenüber dem ConsInd die Rückweisungsoption. Der MatInspector als Suchroutine ist ein sehr schnelles Werkzeug, das in jedem Fall der allgemein üblichen Suche mit IUPAC-Consensi deutlich überlegen ist (Quandt et al, 1995b). Im Laufe der Projektlaufzeit wurden diese Programme ständig erweitert und verbessert. MatInspector 2.0 bietet flexiblere Gestaltung der Ergebnisse und macht dem Benutzer jetzt auch die Matrix-Profile und Referenzen direkt zugänglich. Die Library wurde gleichfalls erweitert (auf mehr als 250 Matrizen, Abb. 3) und eine qualitative Selektion der Matrizen vorgenommen.

MatInd und MatInspector Version 2.0 sind per anonymous ftp von der GSF verfügbar (Adressen siehe oben). Darüber hinaus ist MatInspector über WWW auf den Servern von GSF und GBF nutzbar und an die WWW-Version von TRANSFAC angebunden.

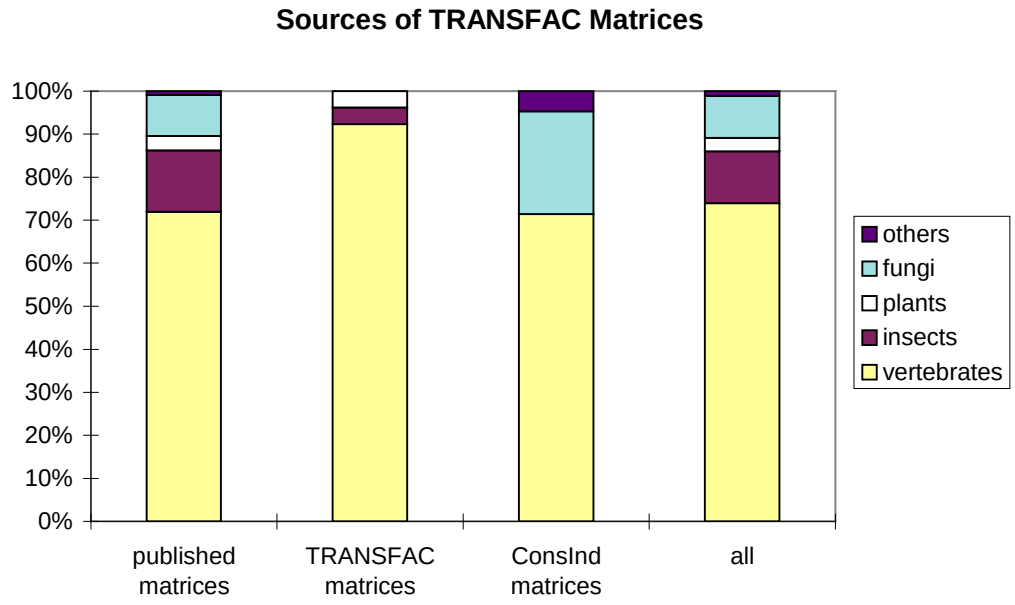


Abb. 3: Verteilung der Nukleotidverteilungs-Matrizen der TRANSFAC-Datenbank nach Quellen und biologischen Gruppen.

3.3 ConsInd/ConsInspector [2]

Aufgrund der teilweise extremen Variabilität regulatorischer DNA-Sequenzelemente muß vermutet werden, daß zusätzliche Sequenzeigenschaften in deren Umgebung einen Beitrag zu ihrer Spezifität liefern. Dies berücksichtigt das an der GSF entwickelte Programm ConsInd, das entsprechende Beschreibungen erstellt, die dann von einem ConsInspector genannten Sequenzanalyse-Programm genutzt werden können, um regulatorische Elemente in beliebig langen DNA Sequenzen zu finden (Frech et al., 1993).

Auf der Grundlage von publizierten Sequenzen bekannter Funktionalität, z. B. dem Bindungsvermögen an einen bekannten Transkriptionsfaktor, werden funktionelle Elemente mit einer Sequenzumgebung definierter Länge, standardmäßig +/- 40 Basenpaare, aus den Sequenz-Datenbanken extrahiert. Z. B. werden, ausgehend von TRANSFAC-Einträgen, die verknüpften Sequenzen aus der EMBL-Datenbank zusammengestellt. Damit wird ein verankertes Alignment durchgeführt, bei dem eine bekannte hochkonservierte Ankersequenz vorgegeben und in einem definierten Sequenzbereich gesucht wird. Ein erschöpfender Satz paarweiser Alignments liefert ein Ähnlichkeitsmaß für alle Sequenzpaare. Daraus wird eine Gewichtung der einzelnen Sequenzen abgeleitet, um zu verhindern, daß sehr ähnliche oder gar identische Sequenzen die

Häufigkeitsverteilung der Basen verfälschen. Mit diesen gewichteten Sequenzen wird ein multiples Alignment erstellt. Das Programm ConsInd errechnet dann für jede Position einen Ci-Wert (Consensus Index), der ein Maß für die Konserviertheit darstellt und im wesentlichen auf der Shannon'schen Informationsdefinition unter Einschluß von Gaps als eigenem Symbol beruht. Der Beitrag jeder einzelnen Sequenz zum Ci-Profil wird durch "random shuffling" außerhalb der hochkonservierten Ankersequenz geprüft. Sequenzen ohne signifikanten Beitrag werden zurückgewiesen.

Es hat sich gezeigt, daß bereits mit sieben bezüglich der oben beschriebenen Gewichtungen unabhängigen Sequenzen Consensus-Beschreibungen erstellt werden können, die für zuverlässige Sequenzsuchen mit dem zugehörigen Analyseprogramm ConsInspector verwendet werden können. ConsInspector sucht in den zu analysierenden Sequenzen zunächst die Ankersequenz, aliniert die der Consensus-Beschreibung entsprechende Sequenzumgebung zu dem Satz der Trainingssequenzen, bewertet die "Paßfähigkeit" und gibt die gefundenen Positionen aus, die über einem vom Nutzer vorgegebenen Schwellenwert liegen.

Es konnte an einer Reihe von Beispielen gezeigt werden, daß die C_i-Werte an den einzelnen Positionen mit deren biologischer Funktionalität kongruent sind (Quandt et al., 1995a). So sind die hochkonservierten Positionen in glucocorticoid-regulierten Elementen (GRE) mit denjenigen Aminosäure-Seitenketten korreliert, die in Kokristallen des Glucocorticoid-Rezeptors mit einer perfekt palindromischen Erkennungssequenz für die DNA-Protein-Kontakte verantwortlich sind (Abb. 4).

ConsInspector ist frei verfügbar und kann von interessierten Nutzern per anonymous ftp (auch über http://www.gsf.de/BIODV) vom ftp-Server der AG BIODV geladen werden (ariane.gsf.de). ConsInspector steht nunmehr in der erheblich erweiterten und verbesserten Version 3.0 zur Verfügung (Frech et al., 1997a). Das Programm wurde durch die Integration einer aus MatInspector abgeleiteten Regionen-Ähnlichkeit erheblich beschleunigt, was für die Analyse sehr langer Sequenzen von großer Bedeutung ist. Außerdem erfolgt nun optional die Doppelstranganalyse und ConsInspector kann die wichtigsten Sequenzformate lesen (EMBL, IG, GCG-Datenbank). Es besteht jetzt auch die Option zur Mehrfachtestung, wodurch die statistische Absicherung der Ergebnisse verbessert werden kann.

Gemeinsam mit dem Teilprojekt der GBF, Braunschweig, haben wir die Library von Consensus-Beschreibungen (erweiterte Gewichtsmatrizen) von 17 auf 37 Einträge mehr als verdoppelt und diese Matrizen wurden ebenfalls im Rahmen dieser Kooperation in die Matrix-Library des MatInspectors übernommen. Die Erstellung dieser Consensi ist relativ arbeitsaufwendig, da die zugrunde liegenden Sites einzeln nachkontrolliert und die erstellten Profile mehrfach validiert werden müssen.

Darüber hinaus wurde ein Programm entwickelt, das es ermöglicht, in nicht-alinierten Sequenzbereichen bekannter gemeinsamer Funktionalität gemeinsame Sequenzmuster aufzufinden (CoreSearch). Die so erstellten Beschreibungen sind identisch mit den aus ConsInd erzeugten Beschreibungen und damit direkt einer ConsInspector-Suche zugänglich (Wolfertstetter et al, 1995).

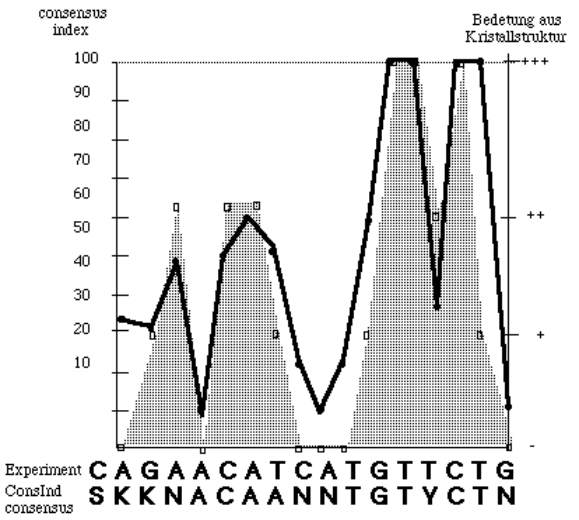


Abb. 4: Übereinstimmung von berechneter und experimentell bestimmter Bedeutung einzelner Nukleotidpositionen im GR-GRE Komplex. Die schwarze Linie stellt die von ConsInd ermittelten Consensus-Indices dar, die graue Fläche entspricht einer aus der Kristallstruktur eines GRE-Inhibitor-Komplexes abgeleiteten Bedeutung einzelner Positionen. - = kein DNA-Kontakt des Glucocorticoid-Rezeptors an dieser Position, + = Zucker-Phosphat-Rückgrat-Kontakt, ++ = unspezifischer Basenkontakt, +++ = spezifischer Basenkontakt. Die experimentell verwendete Sequenz und der von ConsInd ermittelte Consensus sind unterhalb der Profile angegeben.

3.4 TFC (transcription factor cluster analysis tool) [1]

Bei der Einzelelementanalyse mithilfe von PatternSearch, MatInspector oder ConsInspector ist, je nach Stringenz der Suchparameter, u. U. mit einem beträchtlichen Output zu rechnen. Dieser ist in der Regel vom Anwender nachzubearbeiten. Um zu vermeiden, daß jeweils neue Suchläufe gestartet werden müssen, wurde an der GBF ein Datenbanksystem entwickelt, das dem Nutzer zur Verfügung gestellt wird und in das er seinen Output importieren kann. Diese "Results Output Database" (ROD) gestattet nachträglich beliebige weitere Selektions- und Filteroptionen sowie eine damit verbundene Visualisierung (Wingender et al., 1995).

Die in den Output-Listen der genannten Tools bzw. in der ROD enthaltenen vorgeschlagenen TF-Bindungsstellen (TFS) können darüber hinaus auf gehäuftes Auftreten analysiert werden, was mögliche Promotoren oder Enhancer anzeigen könnte. Zu diesem Zweck wurde ein Programm zur probabilistischen Fuzzy-Clusteranalyse (TFC) entwickelt, das für alle TFS einer Sequenz in einem zweidimensionalen Score-/Positionsraum die Zugehörigkeitsgrade zu verschiedenen Clustern errechnet. Fuzzy-Clustering Algorithmen zeichnen sich dadurch aus, daß sie keine ‘harte’ Zuordnung der Daten zu den erkannten Clustern berechnen (Datum gehört

dazu oder nicht), sondern Zugehörigkeitsgrade $\in$ [0,1], die als Wahrscheinlichkeitsverteilung interpretiert werden können. Hierdurch können zum einen Daten, die sich nicht eindeutig einem Cluster zuorden lassen, probabilistisch auf mehrere Cluster aufgeteilt werden, zum anderen Rauscheffekte durch Stördaten i.d.R. besser kompensiert werden. Nach einer erfolgten Clusterung durch den *Fuzzy c-Means* (*FCM*, auch als *Fuzzy ISODATA-Algorithmus* bezeichnet) oder den *Gath & Geva Algorithmus*, können in einer graphischen Benutzeroberfläche qualitativ ausgewählte Cluster auf die DNA-Positionen projiziert und automatisch dokumentiert werden. Das Programm TFC basiert auf einer relationalen Datenbank und enthält diverse Funktionen zum Nachbearbeiten der Daten. Die Ergebnisse von TFC wurden an verschiedenen regulatorischen Regionen wie dem Enhancer des Simian Virus 40 (SV40) verifiziert (s. Abb.5).

Gegenüber bekannten Tools wie PromotorScan (Prestridge, 1994) verfügt TFC somit nicht nur über den Vorteil der fuzzy Datenanalyse, sondern besonders die Zweidimensionalität der Datenanalyse eröffnet einen signifikanten Fortschritt in der Nachbearbeitung der genannten Output-Listen.

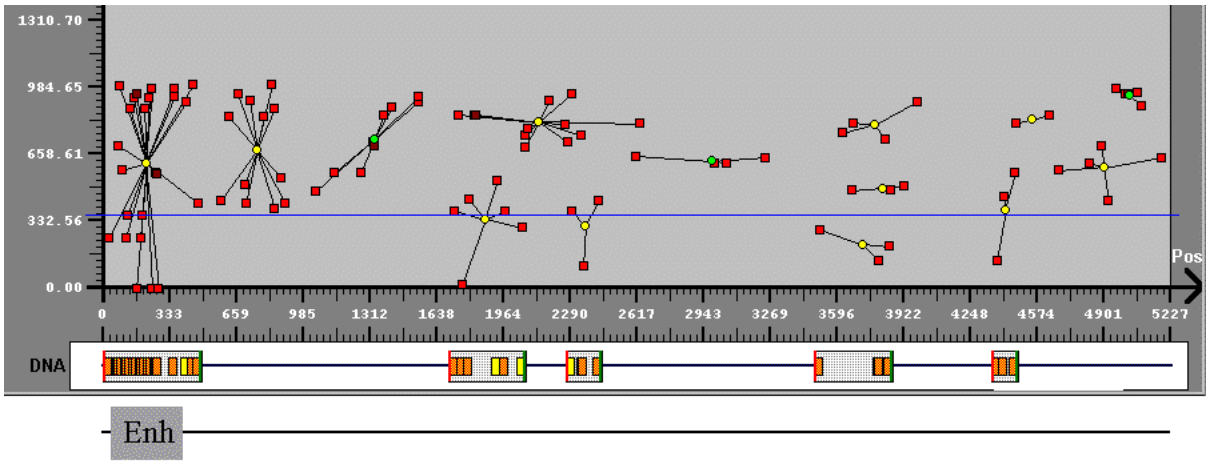


Abb. 5: TFC Cluster-Analyse des SV40 Genoms (5243 bp). Zu erkennen sind eine Reihe Cluster (Sterngraphen) mit den ihnen anhand des höchsten Zugehörigkeitsgrades zugewiesenen Datenpunkten (graue Quadrate). Jeder Datenpunkt enthält mindestens eine potentielle Transkriptionsfaktor-Bindungsstelle. Die horizontale u-Achse gibt die DNA-Position, die vertikale v-Achse einen inversen Score der Daten an. Dies bedeutet eine höhere Qualität der Sites mit kleinerer v-Koordinate. Anhand der interaktiv wählbaren horizontalen Schwellenwertlinie werden diejenigen Cluster auf die DNA projeziert, die mindestens eine Site besitzen, die dieser Schwellenqualität genügt. Der Cluster von besonders hoher Dichte am linken Rand stimmt mit dem experimentell verifizierten SV40 Enhancer (graues ‘Enh’ Rechteck) sehr gut überein.

3.5 GenomeInspector [2]

Für die Anwendung in der Genomforschung besonders wichtig ist die Verfügbarkeit von Werkzeugen, mit denen komplexere Genstrukturen erkannt werden können. Ein besonderes Kennzeichen funktioneller regulatorischer Regionen ist neben der erkennbaren Anhäufung von Bindungsstellen (siehe TFC) eine komplexe Organisation, d.h. bestimmte Abfolge und Abstände der Bindungsstellen auf der DNA-Sequenz. Diese lassen sich auf der Basis von qualitätsabhänigen Abstands-Korrelationen auch vor dem Hintergrund sehr langer Sequenzen gut charakterisieren. Mit GenomeInspector wurde an der GSF ein Programm-Paket entwickelt, das es ermöglicht, interaktiv in Megabasen umfassenden Sequenzen die Korrelationen verschiedener Elementtypen untereinander zu untersuchen und darzustellen. So können z. B. Distanzkorrelationen zwischen offenen Leserastern und funktionellen Merkmalen sehr schnell erkannt werden. Außerdem ist GenomeInspector prinzipiell in der Lage, primäre Suchergebnisse aller zur Erkennung von Sequenzelementen verfügbaren Programme in seine Analyse zu integrieren.

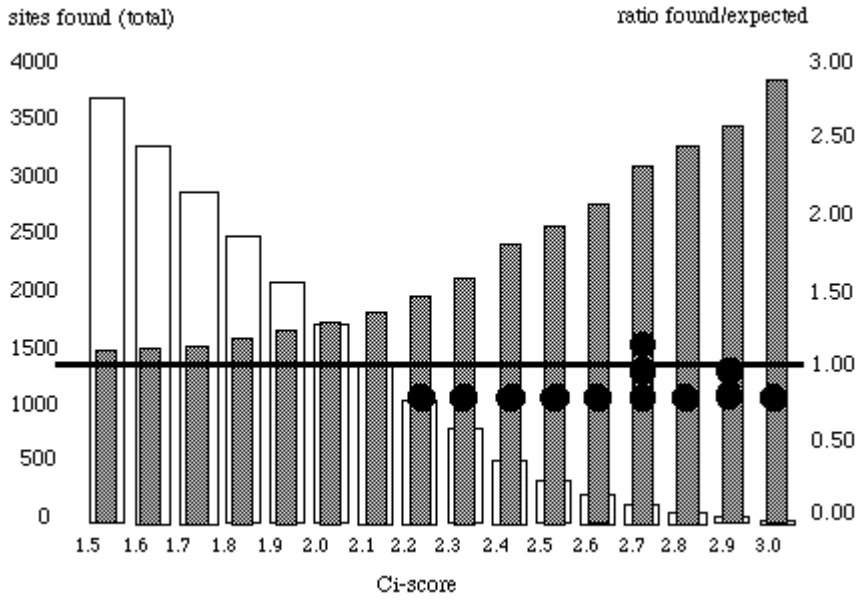


Abb. 6: Anhäufung von potentiellen ABF1 Bindungsstellen mit hohem Ci-Score in der Promoterregion von chromosomalen Hefe-Sequenzen. Die offenen Balken stellen die Gesamtzahlen der mit dem unterhalb der Balken angegebenen Ci-Scores gefundenen ABF1 Stellen dar (linke Skala). Die grauen Balken stellen die Überrepräsentation der ABF1 Stellen im Promoterbereich gegenüber einer errechneten Gleichverteilung dar (rechte Skala). Die schwazen Kreise kennzeichnen die für experimentell verifizierte funktionelle Bindungsstellen ermittelten Ci-scores.

So konnte z. B. gezeigt werden, daß solche potentiellen Bindungsstellen für den Hefe-Transkriptions- und Replikationsfaktor ABF1, die einen besonders hohen Score liefern, in den 500-Basenpaar-Bereichen vor den offenen Leserastern der Hefechromosomen 2, 3, 8 und 11 (ca. 2.4 x 10^6 bp) deutlich überrepräsentiert sind. In diesem Bereich finden sich auch die experimentell bekannten ABF1-Bindungsstellen (Abb. 6). Die Untersuchung höherer Korrelationen wies z. B. für die Klasse der glycolytischen Gene der Hefe eine Kombination aus GCR1-, RAP1- und ABF1-Bindungsstellen als typisches Merkmal aus (Abb. 7).

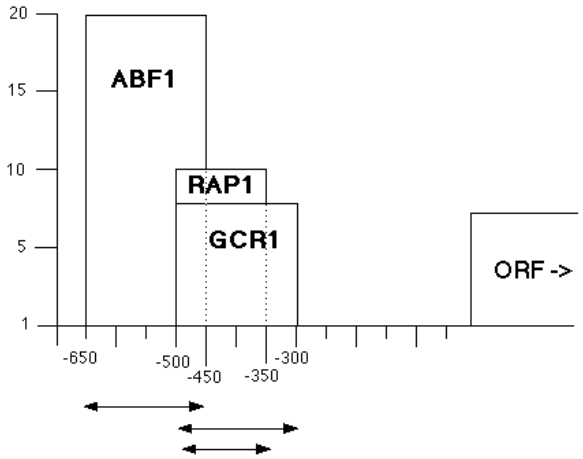


Abb. 7: Charakteristisch überepräsentierte Transkriptionsfaktorbindungsstellen im Promotorbereich von Genen glykolytischer Hefe-Enzyme. Die Skala auf der Y-Achse zeigt, wie oft die entsprechenden Bindungsstellen gegenüber einer errechneten Gleichverteilung überrepräsentiert gefunden wurden (20 = 20-fach überrepräsentiert). Die X-Achse gibt die Nukleotidpositionen vor dem Start des Leserasters (ORF) an. Die Breite der Balken sowie die Doppelpfeile unterhalb der Grafik geben den Bereich an, in dem die zugehörigen Bindungsstellen gefunden wurden.

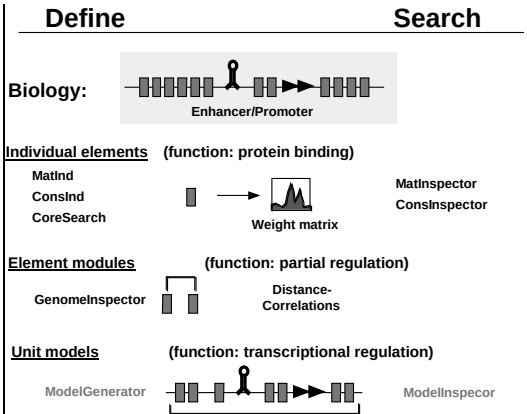
GenomeInspector hat sich nach Beendigung der genomischen Hefe-Sequenzierung in der Praxis als Werkzeug zur Analyse von Gesamt-Genomen bewährt und wurde zu diesem Zweck in seiner Funktionalität ganz erheblich erweitert. Das Grundprogramm erfordert die manuelle Auswahl der zu betrachtenden Partner, um Korrelationen dieser Partner auf der linearen DNA Skala vor dem genomischen Hintergrund zu erkennen (Quandt et al., 1996a, 1996b). Aufgrund der großen Anzahl der bereits verfügbaren Matrizen (MatInspector Library) ist dies in dieser Form nicht mehr erschöpfend zu bewältigen. Wir haben daher den Ablauf automatisiert, so daß GenomeInspector nunmehr in der Lage ist, vollautomatisch aus einem gegebenen Satz von Matrizen diejenigen Paare herauszufiltern, die am besten korreliert sind, wobei wir das bereits definierte Maß für die Überrepräsentation einer Korrelation als Sortier-Kriterium verwenden. Es hat sich gezeigt, daß in diesem Verfahren Anhäufungen (Cluster) von Bindungsstellen besonders berücksichtigt werden müssen, da das Programm ansonsten in erster Linie solche Anhäufungen erkennt. Die entsprechenden Korrekturen erwiesen sich als komplex, konnten aber erfolgreich angeschlossen werden. Es zeigte sich im Rahmen dieser Arbeiten auch, daß GenomeInspector noch erhebliches Potential für weitere Automatisierung aufweist, was aber über den Rahmen des GENUS-Projektes hinausgeht.

3.6 DIALIGN 1.0 [2]

Der bereits beschriebene neue Algorithmus für paarweises und multiples Alignment (Morgenstern et al., 1996) konnte inzwischen als Anwendungsprogramm impelmentiert werden und steht ebenfalls auf unserem ftp-Server zur Verfügung. Es hat sich gezeigt, daß dieses Programm sehr sensitiv Protein-motife aus langen nicht homologen Proteinsequenzen herausfiltern kann, womit sich die Möglichkeit eröffnet, in die Analysen unserer Ebene-2 Verfahren (GenomeInspector, ModelGenerator und ModelInspector) auch Proteinsequenzen miteinzuschließen. Außerdem kann DIALIGN gut verwendet werden um die für ein MatInd Alignment optimalen Bereiche in Sequenzen zu finden, deren Länge weder eine sinnvolle MatInd- noch ConsInd- oder CoreSearch-Analyse erlaubt.

3.7 Vergleich mit dem internationalen Stand [2]

Wir haben in einer Vergleichsstudie alle verfügbaren Programme zur Erkennung von Transkriptionsfaktor-Bindungsstellen gesammelt und anhand eines ausgewogenen Satzes von konkreten publizierten Sequenzbeispielen in ihrer Leistungsfähigkeit miteinander verglichen. Diese Studie wurde von 8 unabhängigen Experten, darunter die Entwickler der Konkurrenzprodukte, geprüft und beurteilt und in zwei Arbeiten publiziert (Frech et al., 1997c; Frech et al., 1997d).



Programs available for detection of individual elements

Definition (binding matrices)	Search programs (with matrix-Library)	Matrix-Libraries	Source
3	2	1	GENUS/BIODV
5	3	1	Rest of the world

Abb. 8: Überblick über die Programmentwicklungen im Teilprojekt GSF und Vergleich mit dem internationalen Stand.

Das Ergebnis zeigt, daß derzeit weltweit 6 Programme zur Erstellung von Matrix-Beschreibungen verfügbar sind, während es 4 Suchprogramme für vorberechnete Matrizen gibt. Je zwei dieser Definitions- und Suchprogramme stammen aus der GSF-Gruppe. Außerdem erwies der Vergleichstest, daß unsere Programme bezüglich der Erkennungsqualität von Einzelelementen den anderen zumindest nicht nachstehen. Bei der Definition multipler Stellen in einem Satz Sequenzen (CoreSearch) und der qualitativen Zuordnung der Vorhersagen mit experimentellen Ergebnissen (ConsInspector und MatInspector) ergaben sich Vorteile für unsere Programme.

Ein entsprechender Vergleich der Ebene-2 Methoden (GenomeInspector) ist aufgrund der unterschiedlichen Leistungsspektren der vorhandenen Programme nicht möglich. Unter den insgesamt fünf bekannten Programmen dieser Kategorie (zwei davon aus unserer Gruppe) gibt es keines, das in ähnlicher Weise wie der GenomeInspector zur Modell-freien Analyse ganzer Genome geeignet ist.

4. Strukturelle Beschreibung regulatorischer Elemente

4.1 Dreidimensionale Struktur der DNA [3]

Signifikante Effekte der Basensequenz auf die räumliche Struktur des DNA-Moleküls wurden in zahlreichen Experimenten mit unterschiedlichen Methoden nachgewiesen. Die mit kristallographischen und NMR-spektroskopischen Methoden aufgeklärten Strukturen von doppelsträngigen Oligonukleotiden und von Protein-DNA-Komplexen zeigen die Variabilität der Konformation des Zucker-Phosphat-Gerüstes in Abhängigkeit von der Basensequenz im atomaren Detail. Globale Strukturmerkmale der Doppelhelix, wie Biegung der Helixachse und die Geometrien der beiden Furchen, können aber auch direkt im Elektronenmikroskop oder indirekt aus Messungen der elektrophoretischen Beweglichkeit sowie Analyse der sich nach enzymatischer Spaltung der DNA ergebenden Fragmente abgeleitet werden. Abb. 9 zeigt an Hand eines Beispiels die von der Sequenz abhängige Variation der Weiten der großen und kleinen Furche sowie die Biegung der Helixachse.

Die daraus folgenden Beziehungen zwischen Basensequenz und strukturellen Parametern der DNA spielen offensichtlich eine wichtige Rolle bei der selektiven Erkennung spezifischer Sequenzen durch regulatorische Proteine. Eine vollständige Beschreibung der funktionellen Eigenschaften regulatorischer Elemente in genomischen DNA-Sequenzen erfordert deshalb, neben der reinen Sequenzinformation auch die durch die Sequenz kodierten

Strukturmodulationen des DNA-Moleküls zu berücksichtigen. Im Hinblick auf die Zielstellung des Verbundprojektes wird diese Forderung durch bekannte Beispiele (TATA-Boxen, C/EBP-Bindungsstellen) unterstrichen, in denen sich strukturelle Merkmale nicht eindeutig auf den Sequenzraum abbilden lassen. Das bedeutet, daß man mit den verfügbaren Algorithmen zum statistischen Vergleich von Sequenzmustern nicht in der Lage ist, experimentell ermittelte Bindungsstellen zu identifizieren.

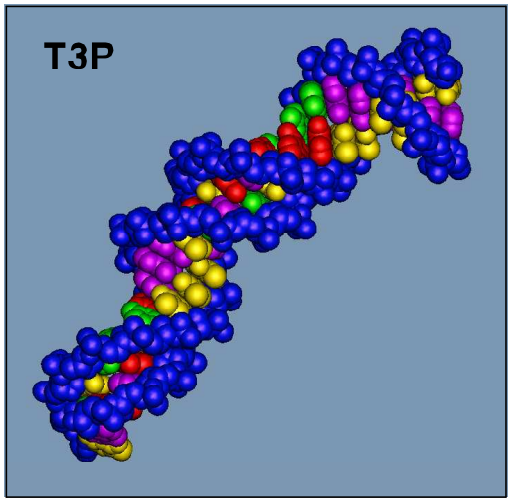


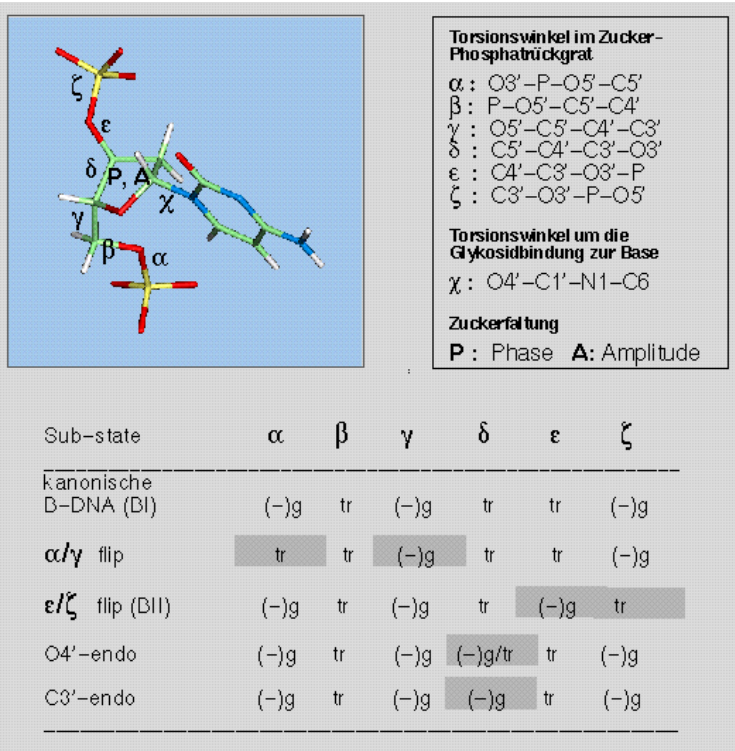
Abb. 9: Räumliches Strukturmodell der DNA-Doppelhelix mit der Basensequenz d(AATTAACCCTCACTAAAGGGAGA) des T3 Promotors. Man erkennt deutlich die von der Sequenz abhängige Variation der Weiten der großen und der kleinen Furche sowie die Biegung der Helixachse.

4.2 Aufbau von Struktur- und Parameterbibliotheken [3]

Um die Suche nach Elementen mit spezifischen Strukturmustern in genomischen DNA-Sequenzen zu ermöglichen, müssen effektive Methoden zur Vorhersage der DNA-Struktur aus der Basensequenz entwickelt werden. Anfängliche Hoffnungen, die Konvertierung beliebiger DNA-Sequenzen in räumliche Strukturen auf Grundlage einer umfassenden experimentellen Datenbasis zu ermöglichen, haben sich bisher nicht erfüllt. Die Schwierigkeiten betreffen insbesondere Probleme der Kristallisation von Oligonukleotiden mit der notwendigen Vielfalt von Sequenzen und die nur eingeschränkten Möglichkeiten, aus NMR-Experimenten vollständige Konformationsdaten für DNA-Moleküle abzuleiten. Außerdem folgt aus der konformationellen Flexibilität der DNA-Struktur eine empfindliche Abhängigkeit von Umgebungseinflüssen, so daß deutliche Unterschiede zwischen Kristall und Lösung beobachtet werden.

Als Alternative zu den gegenwärtig weder in ausreichendem Umfang noch mit der notwendigen Systematik durchführbaren experimentellen Arbeiten bieten sich theoretische Ansätze der molekularen Modellierung an. Im Unterschied zur theoretischen Behandlung des schwierigen Protein- und RNA-Faltungsproblems ist es im Falle der DNA-Struktur

gerechtfertigt, Wechselwirkungen zwischen sequentiell entfernten Teilen des Moleküls zu vernachlässigen. Diese Näherung macht ein solches Projekt rechnerisch durchführbar und erlaubt eine vollständige Konformationsanalyse für lange DNA-Sequenzen, auch unter Berücksichtigung langreichweitiger Korrelationen und dynamischer Aspekte der Struktur.



Sub-state	α	β	γ	δ	ε	ζ
kanonische B-DNA (BI)	(–)g	tr	(–)g	tr	tr	(–)g
α/γ flip	tr	tr	(–)g	tr	tr	(–)g
ε/ζ flip (BII)	(–)g	tr	(–)g	tr	(–)g	tr
O4'-endo	(–)g	tr	(–)g	(–)g/tr	tr	(–)g
C3'-endo	(–)g	tr	(–)g	(–)g	tr	(–)g

Abb. 10: Freiheitsgrade und konformationelle Sub-States des Zucker-Phosphat-Rückgrats der DNA. (-)g, (+)g und tr bezeichnen die jeweils drei möglichen 'staggered' Konformationen gauche(-), gauche(+) und trans.

Am MDC in Berlin-Buch wurden hierzu in enger Zusammenarbeit mit Richard Lavery (Paris) die Programme JUMNA (Junction Minimization of Nucleic Acids) (Lavery et al., 1995) und CURVES (Lavery and Sklenar, 1988, 1989) entwickelt, die heute weltweit in vielen Gruppen für Computersimulationen und theoretische Strukturanalysen helikaler Nukleinsäuremoleküle genutzt werden. Mit den im Hinblick auf die Aufgabenstellung dieses Projektes weiterentwickelten Algorithmen ist es möglich geworden, Effekte der Basensequenz auf die Geometrie der Doppelhelix systematisch zu untersuchen. Solche Studien wurden zunächst für den vollständigen Satz repetitiver Di- und Tetranukleotidstrukturen durchgeführt und haben zu dem Ergebnis geführt, daß jede der Sequenzen durch ein charakteristisches Muster unterschiedlicher Konformationen (Substates) beschrieben werden kann. Das daraus abgeleitete diskrete Substate-Modell der DNA bildet die Grundlage für den Aufbau von Struktur- und Parameterbibliotheken. Die auch mit experimentellen Befunden im Einklang stehenden Substates des Zucker-Phosphat-Gerüstes (Abb. 10) wurden als Startpunkte für Energieminimisierungen und die kombinatorische Suche nach stabilen Konformationen gewählt.

Um aus diesen Ergebnissen eine Strukturbibliothek aufzubauen, müssen die einbezogenen Wechselwirkungen auf nächste oder übernächste Nachbarn beschränkt werden. Diese Näherung entspricht der Annahme, daß die Konformation eines aus einer langen Sequenz herausgegriffenen Nukleotidpaares nur von den Zuständen der 5'- und 3'-flankierenden Nukleotide abhängt. Für den einfachsten Fall (Berücksichtigung nur des jeweiligen Substates mit niedrigster Energie im Modell nächster Nachbarn) ist der entwickelte Algorithmus zum Aufbau der Strukturbibliothek in Abb. 11 schematisch dargestellt. Die Ergebnisse der Konformationsanalye für die sechs unterschiedlichen repetitiven Dinukleotid-Sequenzen bilden den Ausgangspunkt für die Bildung von Startmodellen aller kombinatorisch möglichen repetitiven Tetranukleotid-Sequenzen, die anschließend energetisch optimiert werden. Die strukturelle Paßfähigkeit der aus den optimierten Strukturen extrahierten Strukturfragmente wird durch geometrische 'constraints' (auf die Nukleotide A und B in Abb. 11) während der Energieminimisierung erreicht.

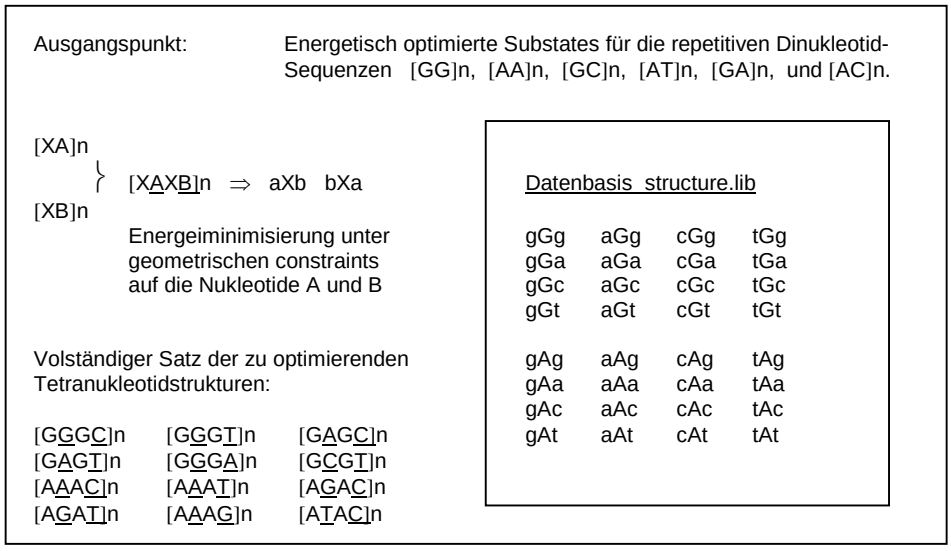


Abb. 11: Schematische Darstellung des Algorithmus zur Generierung der Datenbasis structure.lib im Modell nächster Nachbarn.

Nach dem Aufbau einer vorläufigen Datenbasis für den jeweiligen Zustand niedrigster Energie (single substates) (Knüppel et al., 1996) ist mit dem Abschluß des GENUS-Projektes jetzt auch die vollständige Datenbasis (multiple substates) (Sklenar et al., in Vorbereitung) verfügbar und wird über das world wide web öffentlich zugänglich sein. Mit Hilfe dieser Datenbasis (structure.lib) und der zugehörigen Software können nach Eingabe beliebiger Basensequenzen und praktisch auf Knopfdruck dreidimensionale Modelle der DNA generiert werden. Die Ausgabe im PDB-Format ermöglicht eine anschließende Visualisierung bzw. detailierte Konformationsanalyse mit dem CURVES-Programm. Zusätzlich zu structure.lib

wurde eine speziell auf die Ziele des GENUS-Projektes zugeschnittene zweite Datenbasis (parameter.lib) aufgebaut. Sie enthält einen Satz von Parametern zur Beschreibung struktureller Merkmale der 10 unterschiedlichen Dinukleotidschritte in der Umgebung aller möglichen 5'- und 3'-flankierenden Dinukleotidschritte:

5'- NN XY NN -3'
XY = AA, GG, GA, AG, AT, TA, GC, CG, AC, CA

Zur Berechnung der Parameter wurden mit Hilfe der primären Datenbasis structure.lib die insgesamt 2080 unterschiedlichen Hexanukleotidstrukturen generiert und anschließend mit dem CURVES-Programms analysiert.

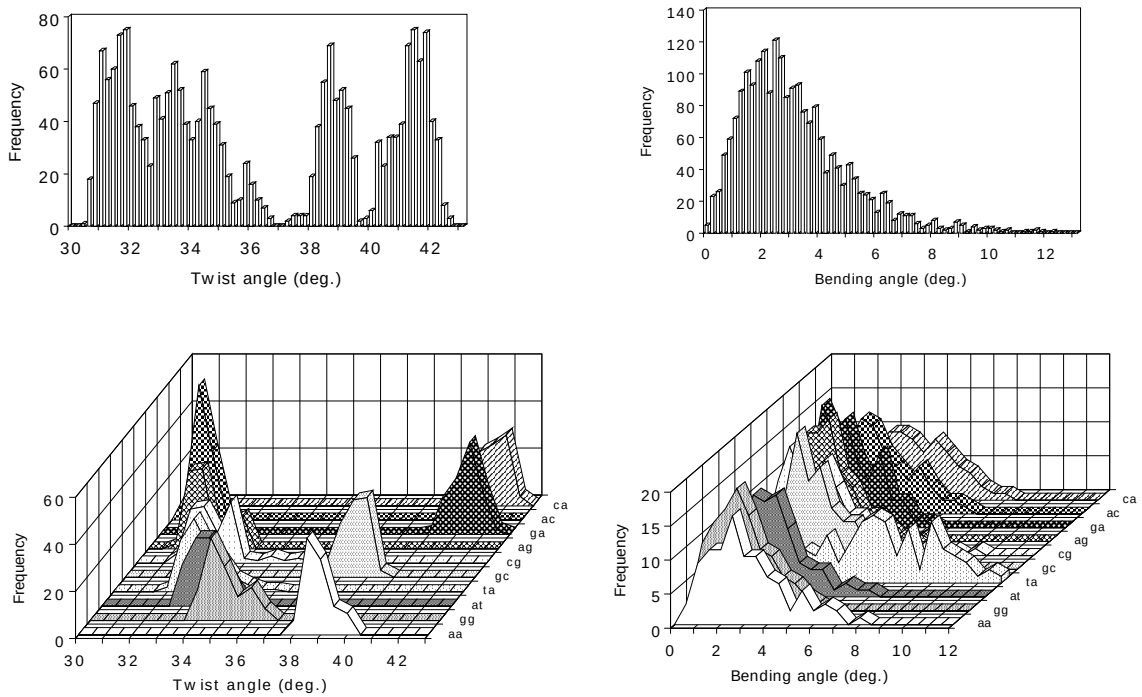


Abb.12: Sequenz-abhängige Verteilung von helikalem Twist und Biegewinkel für den zentralen Dinukleotidschritt in Hexanukleotid-Strukturen. Die Gesamtverteilung (oberer Teil) ist in die Komponenten der 10 verschiedenen Dinukleotidschritte (unterer Teil) zerlegt.

Die gegenwärtige Implementierung umfaßt Werte für die Weiten und Tiefen der großen und kleinen Furche der DNA, für den helikalen Twist und Rise sowie die lokalen Biegewinkel der Helixachse. Diese Auswahl wurde auf Grund erwarteter Korrelationen dieser Parameter mit funktionellen Eigenschaften der DNA getroffen. Die zunächst für single substates generierte Parameter-Bibliothek steht jetzt auch als multiple-substate-Version zur Verfügung und

enthält damit zusätzliche Informationen über die Flexibilität der DNA-Struktur. In Abb. 12 sind die in der single-substate-Version erhalten Verteilungen der helikalen Twist- und Biegewinkel, sowohl für den gesamten Satz als auch für jeden der 10 Dinukleotidschritte separat, als Beispiel dargestellt. Hierdurch werden experimentell beobachtete Sequenzeffekte, zumindest qualitativ, gut widergespiegelt. Im Falle des Twist-Winkels zeigt die Zerlegung der Gesamtverteilung eine deutliche Sequenzabhängigkeit schon auf dem Niveau individueller Dinukleotid-Schritte. Im Unterschied hierzu verweist die breitere Verteilung der Biegewinkel auf eine stärkere Abhängigkeit von den flankierenden Sequenzen. In Übereinstimmung mit experimentellen Befunden wird eine Tendenz zu stärkeren Biegungen in Pyrimidin-Purin-Schritten, besonders ausgeprägt im TA-Schritt, beobachtet.

Die Parameter-Bibliothek ermöglicht eine schnelle Konvertierung von Basen-sequenzen in Struktur-Profile (siehe Abb. 13). Damit sind sowohl systematische Strukturanalysen von Bindungstellen für Transkriptionsfaktoren als auch die Suche nach Regionen mit charakteristischen strukturellen Merkmalen in langen Sequenzen praktisch möglich geworden. In Abb. 14 wird am Beispiel der inneren Kontrollregion des ribosomalen 5S RNA-Gens die Korrelation von Strukturprofilen mit funktionellen Eigenschaften von DNA-Sequenzen demonstriert. Zwei Bereiche der ICR-Sequenz, für die eine Wechselwirkung mit jeweils drei Zinkfinger-Domänen des Transkriptionsfaktors IIIA (TFIIIA) nachgewiesen wurde (Positionen +52 und +78 des 5S RNA-Gens), haben nur sehr geringe Sequenzähnlichkeit, weisen aber fast identische Profile bei der Weite der großen Furche auf.

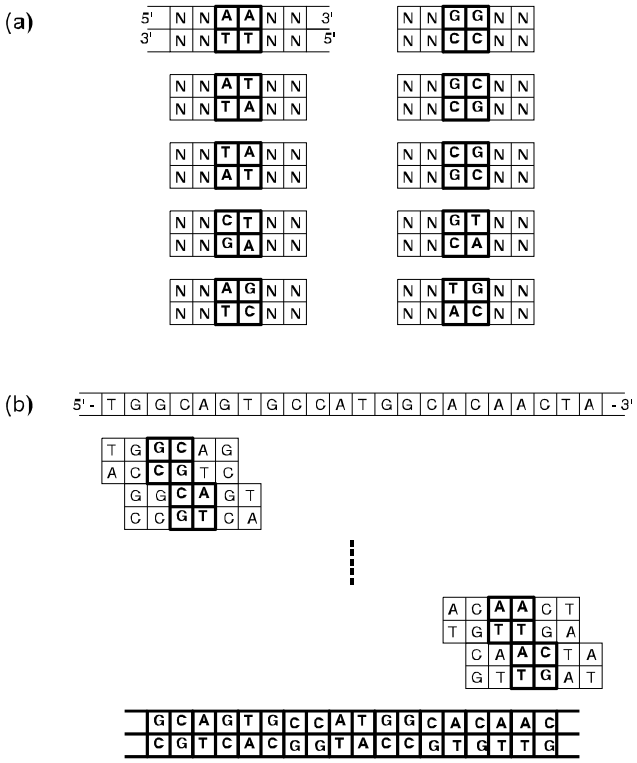


Abb. 13: Hexanukleotid-Library für DNA-Strukturen. (a) Für die zentralen Dinukleotide der generierten Strukturmodelle aller Hexanukleotide werden die relevanten Strukturparameter berechnet und in einer Library niedergelegt. (b) Durch Zusammenfügen überlappender Hexanukleotidstrukturen ergeben sich statische Strukturmodelle für Sequenzen belieber Länge, deren Strukturprofile ohne zusätzliche Berechnungen aus den entsprechenden Einträgen der Library entnommen werden können.

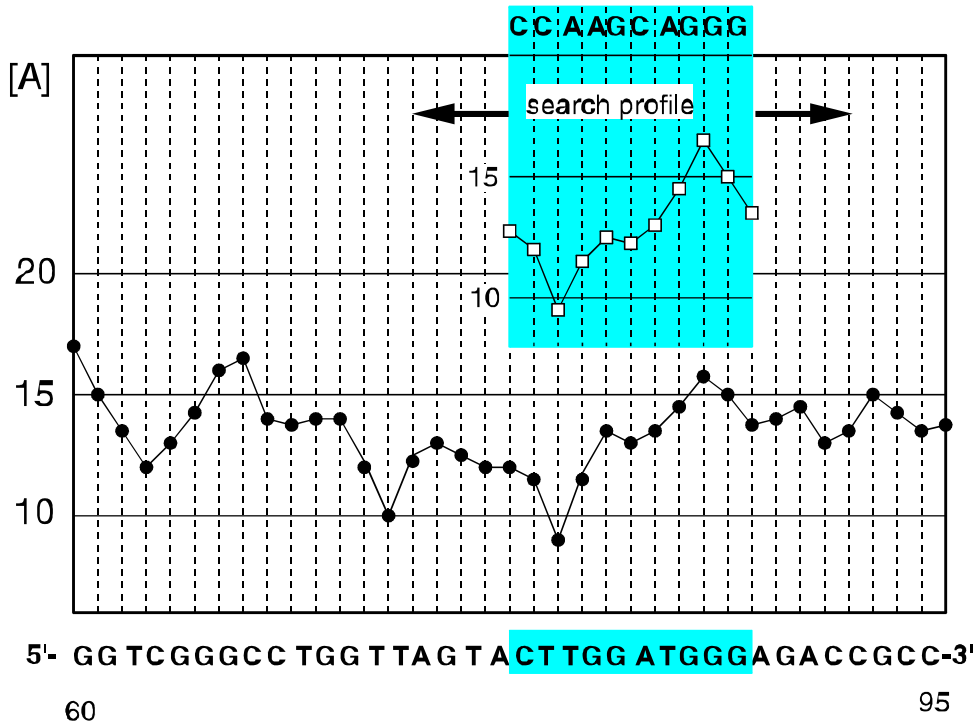


Abb. 14: Verlauf der Weite der großen Furche in der Internen Kontrollregion (ICR) des 5S rRNA-Gens. Gezeigt ist die 5S rRNA-Gensequenz +60/+95, grau schattiert ist ein Element mit experimentell nachgewiesenen Kontakten zu drei Zinkfingermotiven des Transkriptionsfaktors TFIIIA. Das Profil ist nahezu deckungsgleich mit dem einer zweiten Kontaktregion (+52/+61, s. Insert).

4.3 Struktursimulation von TATA-Box-ähnlichen Profilen [1]

Da bekannt ist, daß das TATA-bindende Protein (TBP) in die kleine Furche der DNA bindet, haben wir uns die Frage gestellt, wie konserviert das Profil der Weite der kleinen Furche (minor groove width, mgw) über TATA-Box-enthaltenden Sequenzen tatsächlich ist und ob es eineindeutig mit dieser Sequenz korreliert werden kann. Zur Untersuchung dieser Frage haben wir die strukturellen Eigenschaften von authentischen Hefe-TATA-Boxen mit denen anderer Sequenzen verglichen.

Zunächst wurde dazu die Frage untersucht, wie stark sich der Einfluß der flankierenden Dinukleotide auf die Struktur eines zentralen Hexanukleotids, z. B. einer TATA-Box, auswirkt. Dazu wurden die mgw-Profile sämtlicher Dekanukleotide mit dem Muster NNTATATANN mithilfe der Strukturparameter-Library des MDC berechnet ("Templates"). Das zentrale Hexanukleotid entspricht einer kanonischen Hefe-TATA-Box. Dann wurden die mgw-Profile für sämtliche Dekanukleotide N_{10} berechnet und für jedes der 256 Templates diejenigen Dekanukleotide selektiert, deren mgw-Profile sich von denen der Templates um weniger als einen r.m.s.-Wert von 0.2 unterschieden. Die insgesamt 881 Dekanukleotide, die dieser Bedingung genügten, verteilen sich sehr ungleichmäßig auf die Templates (Tab. 2). Ein besonders selten von anderen Sequenzen zu kopierendes mgw Profil wird dem TATATA-Core offenbar durch flankierende G- bzw. C-Reste aufgeprägt (Tab. 2), wobei besonders ein 5'-flankierendes TG und ein 3'-flankierendes CA ein Template erzeugen, das lediglich genau ein ähnliches Dekanukleotid selektiert, nämlich das selbstähnliche.

Tab. 2: Anzahl der Sequenzen mit zu Templates der Form 5'-NNTATATANN-3' ähnlichen mgw-Profilen

	AN-3'	CN-3'	GN-3'	TN-3'
5'-NA	3319[a]	531	3657	3256
5'-NC	3586	608	3876	3455
5'-NG	574	100	627	531
5'-NT	3339	555	3656	3219

Angegeben ist jeweils die Anzahl der Dekanukleotide, deren mgw-Profil dem den 16 TATATA-Templates ähnlich sind, die von den links angegebenen Dinukleotiden 5'-flankiert und von den oben angegebenen Dinukleotiden 3'-flankiert werden.

Unter den 881 selektierten ähnlichen Dekanukleotiden waren viele, die eine von den Templates völlig verschiedene Sequenz aufwiesen. Zwei Beispielpaare besonders unähnlicher Sequenzen

mit sehr ähnlichen mgw-Profilen waren ATTATATACG und TAACGTTACT, die nur drei gleiche Positionen aufweisen, und CTTATATAAG und TACACGTGTT mit nur zwei identischen Nukleotiden. Dieses Ergebnis zeigt, daß Strukturprofile und Nukleotidsequenzen keine 1:1-Abbildungen voneinander sind, sondern daß die Strukturbeschreibungen tatsächlich unabhängige Kriterien beisteuern können (Karas et al., 1996).

Aus den in Tab. 2 wiedergegebenen Daten geht auch hervor, daß viele der selektierten Sequenzen zu mehreren Templates ähnlich sind. Insgesamt enthalten die 881 selektierten Dekanukleotide nur 25 verschiedene Core-Hexanukleotide. Ihre Verteilung auf die Templates ist in Tab. 3 angegeben. Sequenzen mit dem TATATA-Hexanukleotid im Zentrum werden natürlich von allen 256 Templates selektiert, aber auch ein relativ divergenter Kern wie GTTATA erzeugt noch mgw-Profile, die 192 Templates ähnlich sind.

Tab. 3. Häufigkeitsverteilung der Core-Hexanukleotide unter den Sequenzen mit ähnlichen mgw-Profilen

Hexa-nukleotid	Anzahl der ähnlichen Templates	Hexa-nukleotid	Anzahl der ähnlichen Templates	Hexa-nukleotid	Anzahl der ähnlichen Templates
TATATA	256	ATAACG	120	ACACGT	44
GTTATA	192	CGTTAT	120	CACGTG	42
TATAAC	192	ATATAT	114	CTATAT	20
TAACGT	192	CGTTAA	109	ATATAG	20
ACGTTA	192	AACGTT	84	TTATAG	19
TTATAA	142	AACGTG	58	CTATAA	18
TTAACG	128	CACGTT	53	CTAACG	7
ATATAA	122	GTTAAC	50		
TTATAT	122	ACGTGT	46		

4.4 Anwendung der Strukturanalyse auf TATA-Boxen der Hefe [1]

Mithilfe des ConsInd-Programms der GSF wurden an der GBF Consensus-Beschreibungen für Hefe-TATA-Boxen erstellt. Das Ziel dabei war, statt des Translationsstarts, dem eine 5'-UTR von variabler Länge vorausgeht, und statt des Transkriptionsstarts, der bei den meisten Hefegenen nicht gut definiert ist, einen zuverlässigen Bezugspunkt bei der Lokalisation regulatorischer Elemente von Hefepromotoren zu gewinnen. Mit der errechneten Consensus-Beschreibung wurde die 5'-flankierende Region des *CYC1*-Gens der Hefe untersucht, das zwei

funktionelle und zwei nichtfunktionelle TATA-(ähnliche) Elemente besitzt. Ein funktionelles Element wurde von ConsInspector mit hohem Score (2,34) aufgefunden, das andere hingegen lieferte nur einen sehr geringen Score (1,08), der dem der nichtfunktionellen Boxen glich (1,09). Eine Untersuchung der Furchengeometrien vieler TATA-Boxen ergab jedoch, daß dort stets eine Aufweitung der kleinen Furche auf mindestens 6.5 A flankiert sein muß von einer starken Verengung (Abb. 15). Diese Forderung ist von der zweiten funktionellen TATA-Box erfüllt, nicht jedoch von den beiden nichtfunktionellen TATA-ähnlichen Elementen.

Es zeigt sich somit, daß eine Kombination von u. U. für sich genommen „weichen" Kriterien zu zuverlässigen Resultaten führt. Unter Einsatz des Programms SITEVIDEO eines russischen Kooperationspartners der GBF (Institute of Cytology and Genetics, Novosibirsk), das verschiedene Kriterien automatisch kombiniert und die optimal zwischen einem positiven und einem negativen Trainingssatz diskriminierende Kombination auffindet, gelang es, zwei strukturelle Parameter (Twist und Tiefe der großen Furche) so zu verknüpfen, daß eine ca. 90%-ige Trennung beider Sequenzpopulationen möglich wurde (H. Karas und M. Ponomarenko). Weitere Untersuchungen werden nun darauf gerichtet sein, optimale Kombinationen von strukturellen und statistischen Parametern aufzufinden.

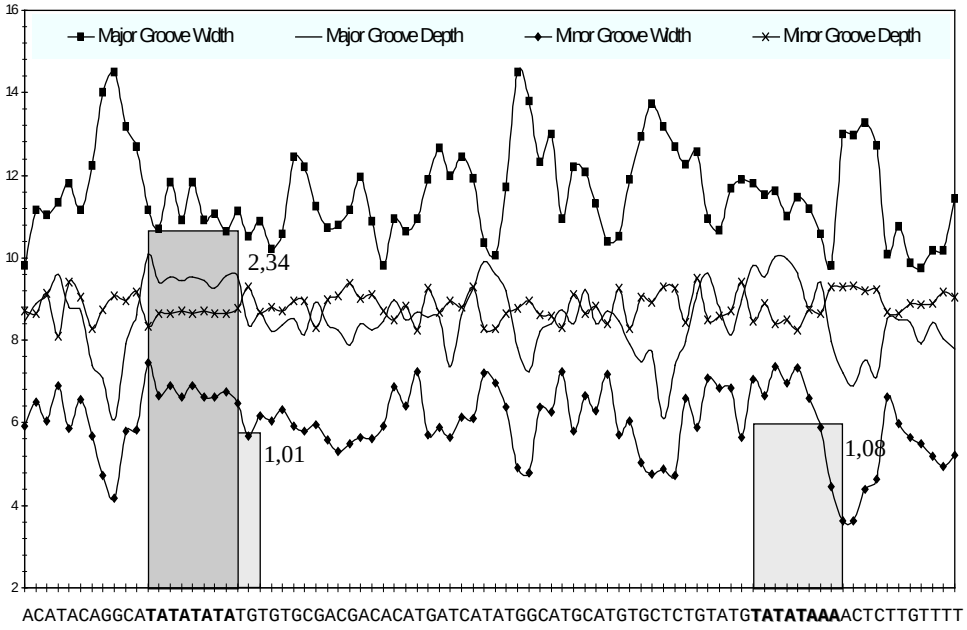


Abb. 15: Profile der Furchenweiten und -tiefen über den funktionellen TATA-Boxen des *CYC1*-Gens der Hefe. Für beide TATA-Boxen sowie für einen um zwei Basen verschobenen Match bei TATA I sind die C_i-scores angegeben, die von ConsInspector errechnet wurden.

4.5 SAGA (structure analysis with genetic algorithms) [1]

Um nicht in allen Fällen alinierter TF-Bindungsstellen erneut den gesamten Suchraum (26 strukturelle Parameter, alle möglichen Fensterpositionen und -größen, ggf. sinnvolle Kombinationen mehrerer Parameter und Fenster) nach indikativen Struktureigenschaften absuchen zu müssen, was offenkundig einen sehr großen Rechenaufwand erfordern würde, wurden an der GBF genetische Algorithmen zur Lösung dieses Problems herangezogen. Mit Hilfe eines solchen Algorithmus wird derjenige Bereich der Sequenzen bestimmt, in dem sie die größte Übereinstimmung bezüglich der gewählten Strukturparameter aufweisen. Die 26 verschiedenen Strukturparameter der Parameterbibliothek aus dem Teilprojekt MDC stehen dem Benutzer zur Verfügung. Für jeden Strukturparameter können weiterhin Strukturprofile erstellt und grafisch dargestellt werden. Die Software wird unter der URL http://www.transfac.gbf-braunschweig.de/cgi-bin/saga/saga.pl zur freien Verwendung über das WWW zur Verfügung gestellt.

Abb. 16: Bezüglich der TATA-Box alinierte DNA-Sequenzen aus der Hefe. Etwa in der Mitte der alinierten Sequenzen ist die TATA-Box zu erkennen. Der von SAGA identifizierte Bereich ähnlicher Weiten der großen Furche (major groove width) etwa 30bp downstream der TATA-Box ist umrahmt dargestellt.

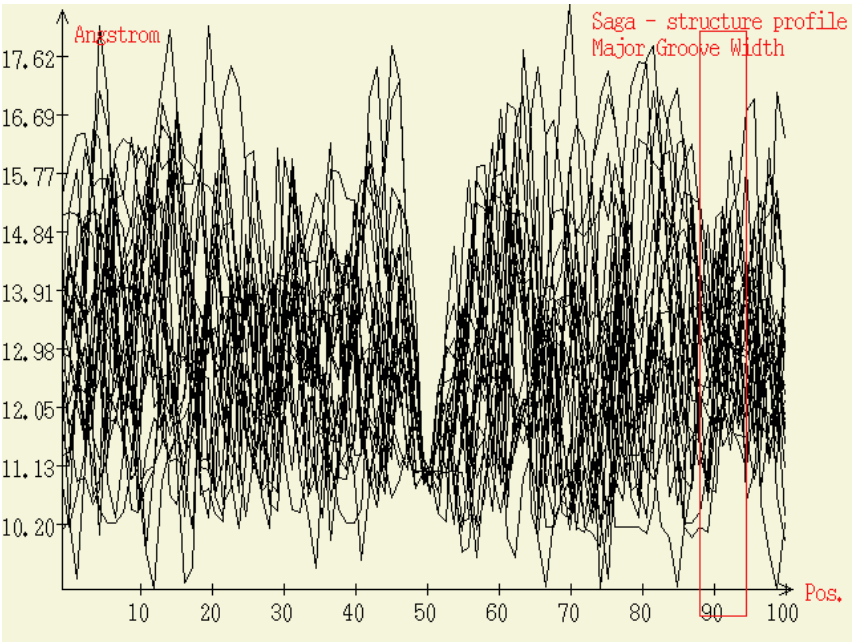


Abb. 17: Profile der Weiten der großen Furche (major groove width profiles) für die in Abb. A dargestellten Sequenzen. Deutlich zu erkennen ist der auch in den Sequenzen konservierte Bereich der TATA-Boxen (etwa in der Mitte). Der von SAGA darüber hinaus identifizierte Bereich ähnlicher Profile ist umrahmt dargestellt.

Im dargestellten Beispiel wurden 25 DNA-Sequenzen aus Hefe-Genen lokal bezüglich der TATA-Box aliniert (Abb. 16). SAGA findet eine strukturelle Ähnlichkeit in der Weite der großen Furche 30bp downstream der TATA-Box (Abb. 17), also einer Position, die dem Transkriptionsstartpunkt entspricht. Diese Ähnlichkeit ist aus der reinen Basensequenz nicht ersichtlich (Abb. 16).

4.6 Einbindung struktureller Parameter in die Sequenzanalyse [2]

Eine direkte Einbindung struktureller Parameter in die ConsInd-Analyse erwies sich derzeit als nicht gangbar, weshalb alternative Lösungen vorbereitet wurden, die auf einer kombinatorischen Auswertung getrennter Analysen beruhen. Wir haben die Vorarbeiten soweit abgeschlossen, daß strukturelle Analysen sehr rasch in die Bewertung der Matrix-Matche miteinbezogen werden können, sobald entsprechende Tools zur Verfügung stehen, womit aufgrund der im Teilprojekt MDC erfolgreich abgeschlossenen Generierung der Hexanukleotid-Library (Karas et al., 1996) in Kürze zu rechnen ist.

5. Positionskorrelationen bei regulatorischen Elementen [4, 1]

Viele regulatorische Elemente haben palindromischen Charakter. Es schien daher denkbar, daß Variationen in der einen Hälfte der Erkennungssequenz von einer komplementären Veränderung in der anderen Hälfte begleitet werden, was die beträchtliche Variabilität der bekannten Bindungssequenzen mancher Transkriptionsfaktoren wie NF-1 erklären helfen könnte.

Um solche Korrelationen aufzuspüren, wurde an der Universität Bielefeld ein Programm erstellt, das für je zwei Positionen aus einem Satz alinierter Sequenzen die Korrelation zwischen den dort zu findenden Eintragungen gemäß einer von der Shannon'schen Informationstheorie abgeleiteten Formel berechnet:

$$cor(i, j) = cor(i, j; S) = - \sum p(i, j; x, y) \log(p(i; x)\, p(j; y) / p(i, j; x, y)) \qquad (x, y \text{ in } A)$$

wobei

S eine Familie alinierter Sequenzen über einem Alphabet A ist (z. B. über der Menge der Nukleotide oder der Menge der Aminosäuren),

p(i;x) die relative Häufigkeit des Vorkommens von x an Position i in den Sequenzen aus S und

p(i,j;x,y) die relative Häufigkeit des Vorkommens von x an Position i und gleichzeitig von y an Position j in den Sequenzen aus S

bezeichnet.

Konstante Positionen (d.h. solche der Entropie 0, also ohne Informationsgehalt) besitzen gemäß dieser Formel keinerlei Korrelation mit gleich welchen anderen Positionen, hochkorreliert sind dagegen hochmutable (also hochentropische bzw. hochinformative) Positionen, an denen jedoch nur vergleichsweise wenige Nukleotid- oder Aminosäurekombinationen auftreten.

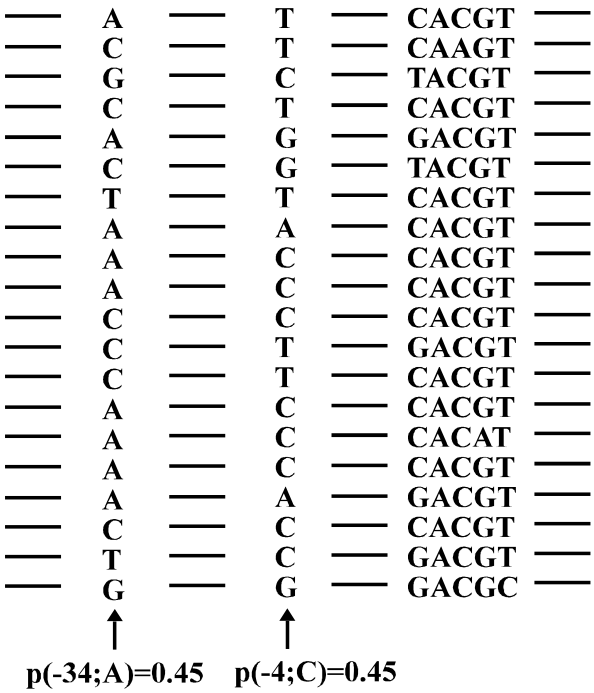


Abb. 18: Positionskorrelation in der Nähe von Bindungsstellen für pflanzliche G-Box-bindende Proteine (GBP). Das Beispiel illustriert die Korrelation der Positionen -34 und -4 (+1: erstes Nukleotid der konservierten Core-Sequenz CACGT). Signifikant über dem Erwartungswert liegen das Auftreten von C bzw. T bei -4, wenn bei -34 ein A bzw. T auftritt. Für das erstgenannte Paar ist das gefundene Vorkommen 0,25 gegenüber einer *a priori* Wahrscheinlichkeit von 0,20.

Untersucht wurden u. a. kompilierte Bindungsstellen der Transkriptionsfaktoren NF-1, C/EBP, CREB. In keinem Fall konnten innerhalb der eigentlichen Bindungssequenzen Positionskorrelationen aufgefunden werden. Zunächst an C/EBP-Sites, dann an Erkennungsstellen der pflanzlichen G-Box-bindenden Proteine (GBP) wurden jedoch derartige Korrelationen in der weiteren Sequenzumgebung gefunden (Abb. 18). Zur Kontrolle wurden gleichartige Analysen an zeilen- und ebenso an spaltenweise randomisierten Sequenzen vorgenommen (spaltenweise Randomisierung erhält die Qualität des Alignments sowie die Entropie der einzelnen Positionen). Dabei zeigte sich, daß die Anzahl hochkorrelierter Positionspaare nur unwesentlich abnahm. Wurden dagegen „Cliquen“ von paarweise hochkorrelierten Positionen betrachtet, so nahm deren Anzahl in den randomisierten gegenüber den ursprünglichen Sequenzen deutlich ab, was mit steigender Cliquengröße zunehmend auffälliger wurde (Abb. 19).

Hochkorrelierte Positionen wurden in einem Bereich von 200 Basenpaaren um definierte Regulationselemente herum gefunden. Man könnte also eine Verbindung mit der Nukleosomen-Assoziation oder höheren Ebenen der Chromatin-Organisation vermuten. Eine gesicherte biologische Bedeutung dieses Befundes gibt es jedoch noch nicht. Ähnliche Befunde ergaben sich bei der Analyse alinierter C/EBP-Bindungsstellen, während an alinierten CREB-Sites derartiges nicht beobachtet werden konnte.

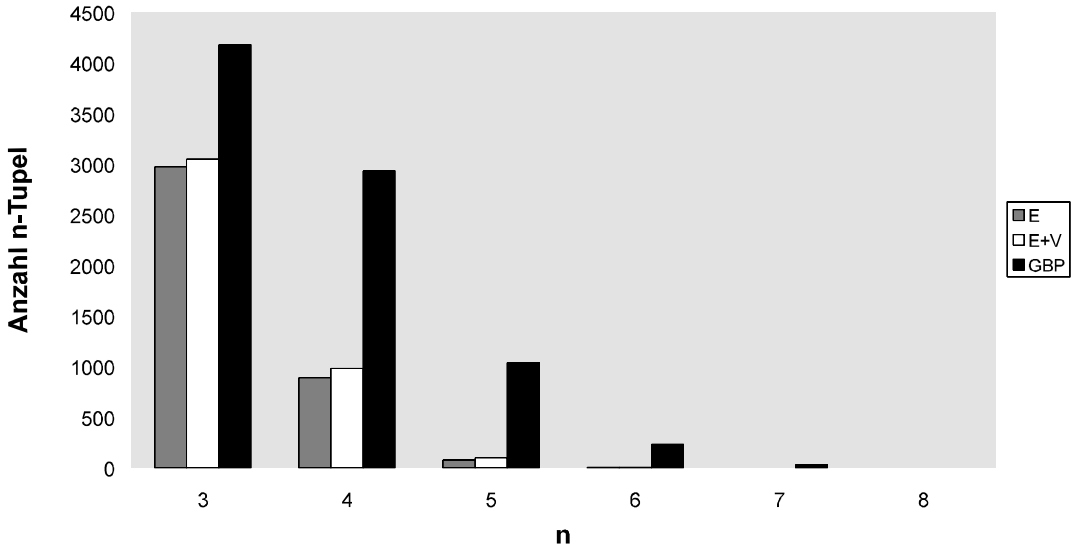


Abb. 19: Anzahl hochkorrelierter *n*-Cliquen (in den 10% höchstkorrelierten Positionspaaren) in der Umgebung alinierter GBP-Bindungsstellen.

Es besteht folglich nun die Aufgabe, diesen überraschenden Befund an weiteren Sätzen von Bindungsstellen zu verifizieren und in die vorhandenen Suchroutinen einzuarbeiten. Ferner bleibt zu prüfen, inwieweit sich die Methode auch zur Untersuchung der Korrelation von reellwertigen, sequenz-abhängigen Parametern eignet, z. B. zur Korrelationsanalyse von Furchenweiten und -tiefen.

Über die Verifizierung regulatorischer Elemente hinaus kann die Korrelationsanalyse dazu verwandt werden, zwischen alternativen Alignments zu entscheiden, da das bessere Alignment im Zweifel dasjenige sein sollte, welches die höheren Korrelationen aufweist. Ferner kann sie zur Berechnung eines auf Positionskorrelationen beruhenden Ähnlichkeitsmaßes genutzt und so z. B. für die Analyse von Phylogenien herangezogen werden (Dress, Terhalle, 1997).

In Kooperation mit Partnern außerhalb des GENUS-Verbundes wurden die Korrelationsmethoden zusammen mit in Bielefeld entwickelten Clusterverfahren auch in anderen Zusammenhängen angewandt, so zur Analyse von Bandenmustern zwecks Bestimmung von Taxonomie und Phylogenie von *Guinea Yam* (Ramser et al.) und seit kurzem auch zur Analyse von Kombinationen fluoreszenzmarkierter Oberflächenproteine an Blutzellen von ALS-Patienten (ALS = a*myotrophe laterale Sklerose*).

Die Anwendung der Korrelationsanalysen auf Proteinstrukturen, speziell auf die DNA-Bindungsdomänen (DBD) von Transkriptionsfaktoren, erwies sich als sehr vielversprechend.

Auf bZIP-Domänen angewandt ergaben sich deutliche Korrelationen zwischen Positionen, die in der basischen, DNA-kontaktierenden Region liegen, und solchen, die sich in der “leucine zipper” (ZIP)-Dimerisierungsdomäne befinden. Dieser Befund zeigt, daß DNA-Bindungsspezifität und Präferenz für bestimmte Dimerisierungspartner dieser Transkriptionsregulatoren miteinander verknüpft sind.

Derartige Korrelationseigenschaften ließen sich auch mithilfe der Splits-Analyse zur Erkennung von Verwandtschaftsbeziehungen nutzen, wobei sich für die bZIP-Domänen allerdings kein anderes Bild ergab, als bei der Verwendung allgemein verbreiteter Alignment- und Phylogenieprogramme. Mit solchen war es jedoch schwierig, eindeutige Verwandtschaftsbeziehungen für eine andere Klasse von Transkriptionsfaktoren aufzuzeigen, nämlich für die der bHLH-Proteine (bHLH: basic region / helix-loop-helix domain). Diese Proteine sind für die Kontrolle der Entwicklung eines Organismus, insbesondere bei Zelldifferenzierungsereignissen, von besonderer Bedeutung.

Da die Loop-Region dieser Domänen variable Längen besitzt, wurde deren Sequenz zur Vereinfachung der Analyse durch ihre Länge ersetzt. Es stellten sich bei der Analyse von 86 bHLH-Domänen acht “Positionen” als signifikant korreliert heraus, davon drei in der basischen (DNA-kontaktierenden) Region, zwei in Helix 1 der Dimerisierungsdomäne (HLH), die Looplänge, sowie zwei Positionen in Helix 2. Reduziert man die bHLH-Domänen auf diese acht Merkmale und unterzieht sie Verwandtschaftsbetrachtungen, ergibt sich das in Abb. 20 gezeigte Bild.

Im Rahmen der zusatzfinanzierten Verlängerung sind in Bielefeld in Zusammenarbeit mit Prof. Dr. Atchley von der North Carolina State University diese Korrelationsanalysen an bHLH-Proteeinen vertieft worden. Hierbei wurden verstärkt auch Korrelationen zwischen Positionen einerseits und Zugehörigkeit zu funktional definierten Subfamilien andererseits berücksichtigt (Atchley et al.). Die Resultate letzterer Korrelationsanalyse konnten erfolgreich dazu verwandt werden, eine Reihe von bHLH-Proteinen, die nicht im untersuchten Datensatz vertreten waren, diesen Subfamilien zuzuordnen, also zu klassifizieren.

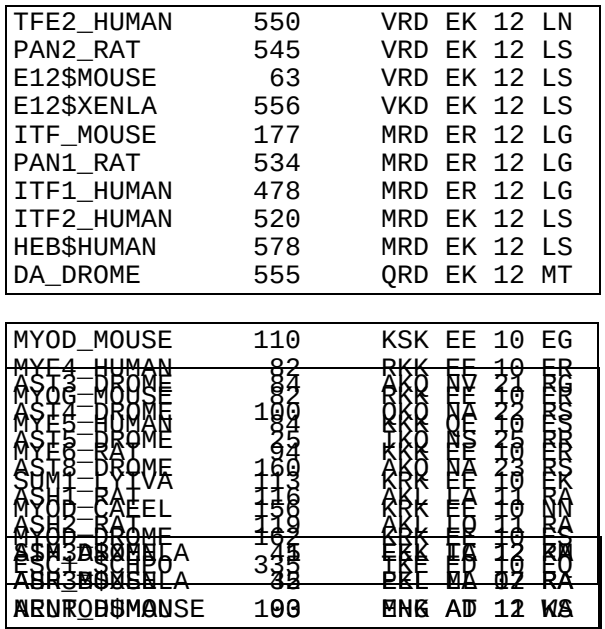


Abb. 20: Die aufgrund einer Positionskorrelationsanalyse vorgenommene Klassifizierung von basic - helix-loop-helix -Proteinen (bHLH). Nicht gezeigt sind die Ergebnisse für die bHLH-ZIP-Faktoren, die zusätzlich zur bHLH-Domäne noch einen leucine zipper als weiteres Dimerisierungsinterface besitzen und somit strukturell wie funktionell eine eigene Familie bilden. Die Bezeichnungen am linken Spaltenrand sind SwissProt ID bzw. (mit einem $-Zeichen zwischen Protein- und Speziesbezeichnung) ID aus eigenen konzeptionellen Translationen entsprechender DNA-Sequenzen; die darauf folgende Zahl gibt den Beginn der bHLH-Domäne im Proteinmolekül an. Weiterhin sind die korreliert gefundenen Aminosäurereste der bHLH-Positionen 2, 13, 14 (basische Region), 17, 20 (Helix 1), 39, 40 (Helix 2), sowie, zwischen Helix 1 und 2, die Looplänge angegeben. Zumindest in einigen Fällen (erster und zweiter Kasten) entspricht die Einteilung auch funktionellen Gesichtspunkten.

```
LIN32$CAEEL      2     SNT VD 11 EC
ATO$DROME      256     LQN QD 11 SA
SCL_HUMAN      188     IQN GA 11 NF
TAL2$HUMAN       3     IQN SA 11 NF
TAL2$MOUSE       3     IQS NA 11 NF
LYL1_HUMAN     138     VQN GA 11 GF
LYL1_MOUSE     150     VQH GA 11 GF
TWST_DROME     363     VQS DK 10 DF
TWST_MOUSE     113     VQS EA 10 DF
TWST_XENLA      73     VQS ES 10 DF
EC2$HUMAN       73     QQS TT 11 AH
TH1$MOUSE       95     GES SA 11 AY
HEN1_HUMAN      76     TEA LA 11 SY
HEN1_MOUSE      76     TEA LA 11 SY
HEN2_HUMAN      78     SEA LA 11 SY
SGC1$YEAST     181     QIN TA 44 LY

HAIR_DROME      32     SAR NN 16 QE
ESM5_DROME      19     VAR KD 14 RK
ESM7_DROME      14     VAR KD 14 RK
ESM8_DROME      11     VAR KD 14 RQ
DPN$DROME       41     TAR HN 16 QS
HES1_MOUSE      35     SAR ES 16 RN
HES2_RAT        14     SAR ES 16 RE
HES5_RAT        17     LDR SE 15 KH
PHO4_YEAST     251     ENR VH 15 RH
NUC1_NEUCR     668     TNR SQ 49 KQ
ARLC_MAIZE     413     KEK EL 08 KE

DELI$DROME      95     KRE TE 18 TM

CBF1_YEAST     223     DEN TN 07 QK

INO2_YEAST     237     WIN EE 13 KS

INO4_YEAST      46     IEL AD 10 LS
```

6. Status des Verbundprojektes

In Tab. 4 sind die im Antrag festgelegten Ziele des GENUS-Projektes den bisher erreichten Ergebnissen gegenübergestellt. Die bisherigen Entwicklungen aus dem Verbund haben bereits zu einer Reihe von *peer-reviewed* Publikationen geführt (34 Arbeiten), 12 weitere Publikationen sind durch Anwendungen in Zusammenarbeit mit externen Kooperationspartnern entstanden.

Tab. 4: Ziele und Ergebnisse des GENUS-Verbundprojektes.

Ziele	*Ergebnisse*
Datenbank genregulatorischer Elemente und Faktoren	TRANSFAC
funktionelle Klassifikation	TRANSFAC-Tabellen GENE, CLASS
statistische Beschreibung	MatInd, ConsInd, Korrelationsanalyse, TFC
strukturelle Beschreibung	Hexanukleotid-Library
Suchprogramme	MatInspector, ConsInspector, GenomeInspector
Beziehungen zwischen Statistik und Struktur	SAGA
Einarbeitung der Ergebnisse in die Datenbank	TRANSFAC-Tabellen MATRIX, CONS
Software-Integration	WWW-Server bei GBF und GSF

Insgesamt hat das Verbundprojekt schon jetzt eine Reihe von Computer-Tools hervorgebracht, die schon einzeln, erst recht aber in ihrer Kombination (Datenbank / Statistik / Struktur) im internationalen Vergleich einmalig sind. Sie stellen ein wichtiges Werkzeug für die Genomforschung dar, da sie für die Abschätzung regulatorischer Potentiale von Genomabschnitten unentbehrlich sein werden. Ihre Anwendung im Rahmen von Genomforschungsprojekten wird zu ihrer ständigen Verfeinerung und Steigerung ihrer Leistungsfähigkeit führen.

Dies zeigt, dass das GENUS-Verbundprojekt die gestellten Aufgaben erfüllt hat und in vielen Punkten Ergebnisse erzielen konnte, die über die ursprüngliche Planung deutlich hinausgehen.

Durch die bisherigen Arbeiten haben sich aber auch eine Reihe von wichtigen Fragestellungen ergeben, die im Rahmen dieses Projektes nicht mehr bearbeitet werden können, und die auch über die reine Anwendung hinausgehen. Dazu gehören:

- Wie ist der Einfluß der Proteinkomponente auf die Struktur eines Regulationselementes bei der Wechselwirkung zu bewerten? Kann für diese Wechselwirkungen eine Art "regulatorischer Code" definiert werden?
- Wie sieht die Syntax komplexer Regulationseinheiten wie ganzer Promotor- oder Enhancerregionen und deren Wechselspiel mit LCR (lokalen Kontrollregionen) und S/MARs (scaffold/matrix attachment regions) aus?
- Wie kann der Einfluß regulatorischer Mechanismen z. B. der Signaltransduktion auf die Genregulation beschrieben und modelliert werden?

- Wie lassen sich komplexe zelluläre Vorgänge wie morphogenetische Prozesse und Zelldifferenzierung auf der Grundlage molekularer Mechanismen der Genregulation modellieren?

7. Veröffentlichungen

Publikationen über GENUS-Entwicklungen:

Bandelt, H.-J. and Dress, A. W. M. (1994). An Order Theoretic Framework for Overlapping Clustering. Discrete Mathematics 136, 21-37.

Bandelt, H.-J. and Dress, A. W. M. (1993) A Relational Approach to Split Decomposition, Journal of Classification; Universität Bielefeld, FSP Mathematisierung-Strukturbildungsprozesse, Materialien LXVIII (1993).

Butzlaff, M., Dahmen, W., Diekmann, S., Dress, A. W. M., Schmitt E. and von Kitzing, E. (1994). A hierarchical approach to force field calculations through spline approximations. Journal of Mathematical Chemistry 15, 77-92.

Deutsch, A., Dress, A. W. M. and Rensing, L. (1993). Formation of morphological differentiation patterns in the ascomycete Neurspora crassa. Mechanisms of Development 44, 17-31.

Dress, A. W. M. (1995). Some Mathematical Problems Arising in Molecular Bioinformatics. In: C. J. Colbourn and E. S. Mahmoodian (eds.), Combinatoric Advances, 91-109.

Dress, A. W. M., Moulton, V. L. and Terhalle, W. F. (1996). T-Theory - An overview. Eur. J. Combinatorics, 17, 161-175.

Dress, A. W. M. and Terhalle, W. F. (1995). Rewarding maps - on greedy optimization of set functions. Adv. Appl. Math., 16, 464-483.

Dress, A. W. M., Mueller, A. and Orville-Thomas, W. J. (eds.) (1995). Topological Aspects of Molecular Structures. Special Issue, Journal of Molecular Structure, Theochem, Theory and Modelling in Chemistry 336/2-3.

Dress, A. and Terhalle, W. (1997). Similarities of Aligned Sequences Based on Correlations of Positions. In preparation.

Frech, K., Brack-Werner, R. and Werner, T. (1996). Common modular structure of *Lentivirus* LTRs. Virology 224, 256-267.

Frech, K., Dietze, P., Werner, T. (1997a) ConsInspector 3.0: new library and enhanced functionality. Comput. Appl. Biosci. 13, 109-110.

Frech, K. and Werner, T. (1997b) Specific modeling of regulatory units in DNA-sequences, Pacific Symposium on Biocomputing 97 (Hrsg. R. Altman, A. K. Dunker, L. Hunter, T. E. Klein), 151-162.

Frech, K., Quandt, K. and Werner, T. (1997c). Software for the analysis of DNA sequence elements of transcription. Comput. Appl. Biosci. 13, 89-97.

Frech, K., Quandt, K. and Werner, T. (1997d). Finding protein binding sites in DNA sequences - the next generation. Trends in Biochem. Sci., 22, 103-104.

Frech, K., Quandt, K. and Werner, T. (1997e). A new method to develop highly specific models for regulatory DNA regions. Lecture Notes in Computer Sciences (LNCS), Springer Verlag, im Druck.

Karas, H., Knüppel, R., Schulz, W., Sklenar, H. and Wingender, E. (1996). Combining structural analysis of DNA with search routines for detection of transcription regulatory elements. Comput. Appl. Biosci. 12, 441-446 (1996).

Kel, A., Kel, O., Ischenko, I., Kolchanov, N., Karas, H., Wingender, E. and Sklenar, H. (1996). TRRD and COMPEL databases on transcription linked to TRANSFAC as tools for analysis and recognition of regulatory sequences. Computer Science and Biology. Proceedings of the German Conference on Bioinformatics (GCB'96), R. Hofestädt, T. Lengauer, M. Löffler, D. Schomburg (eds.). University of Leipzig, Leipzig 1996, pp. 113-117.

Kel, O. V., Romaschenko, A. G., Kel, A. E., Wingender, E. and Kolchanov, N. A. (1995). A compilation of composite regulatory elements affecting gene transcription in vertebrates. Nucleic Acids Res. 23, 4097-4103.

Knüppel, R., Dietze, P., Lehnberg, W., Frech, K. and Wingender, E. (1994). TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J. Comput. Biol. 1, 191-198.

Lavery, R., Zakrzewska, K. and Sklenar, H. (1995). JUMNA (junction minimization of nucleic acid structures). Comp. Phys. Comm., 91, 135-158.

Leijon, M., Zdunek, J., Fritzsche, H., Sklenar, H. and Grdslund, A. (1995) NMR studies and restrained molecular-dynamics calculations of a long A+T stretch in DNA: effects of phosphate charge and solvent approximations. Eur. J. Biochem., 234, 832-842.

Morgenstern, B., Dress, A. and Werner, T. (1996). Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. 93, 12098-12103.

Quandt, K., Frech, K., Herrmann, K. and Werner, T. (1995a). A consensus match scoring system that is correlated with biological functionality, in Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism (Eds. D. Schomburg, U. Lessel), 47-57.

Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995b). MatInd and MatInspector - New fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 23, 4878-4884.

Quandt, K., Grote, K. and Werner, T. (1996a). GenomeInspector: Basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. Genomics 33, 301-304.

Quandt, K., Grote, K. and Werner, T. (1996b) GenomeInspector: A new approach to detect correlation patterns of elements on genomic sequences, Comput. Appl. Biosci. 12, 405-413.

Quandt, K. Frech, K. and Werner, T. (1997) Analysis of transcription control regions by detection and spatial organization of individual transcription factor binding sites. Mol. Biol., in press.

Wingender, E. (1994) Recognition of regulatory regions in genomic sequences. J. Biotechnol. 35, 273-280.

Wingender, E., Dress, A., Sklenar, H., Werner, T. (1995). Verbundprojekt GENUS Vergleichende Analyse und Erkennung genregulatorischer Nukleinsäure-Sequenzen, Statusseminar des BMBF BIOINFORMATIK (Hrsg. Projektträger Informationstechnik des BMBF bei der DLR e.V., G. Wolf, R. Schmidt, M. van der Meer), 105-125.

Wingender, E., Dietze, P., Karas, H. and Knüppel, R. (1996a). TRANSFAC: A database on transcription factors and their DNA binding sites. Nucleic Acids Res. 24, 238-241.

Wingender, E., Karas, H. and Knüppel, R. (1996b). TRANSFAC Database as a Bridge between Sequence Data Libraries and Biological Function. Pacific Symposium on Biocomputing '97 (PSB'97), R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein (eds.). World Scientific, Singapore - New Jersey - London - Hong Kong 1996, pp. 477-485.

Wingender, E. (1997). Classification scheme of eukaryotic transcription factors. Molekularnaya Biologiya, in press.

Wingender, E., Kel, A. E., Kel, O. V., Karas, H., Heinemeyer, T., Dietze, P., Knüppel, R., Romaschenko, A. G. and Kolchanov, N. A. (1997). TRANSFAC, TRRD and COMPEL: Towards a federated database system on transcriptional regulation. Nucleic Acids Res. 25, 265-268.

Wolfertstetter, F., Frech, K., Herrmann, G. and Werner, T. (1995). Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. Comput. Appl. Biosci. 12, 71-80.

Weitere Publikationen über GENUS-Anwendungen:

Atchley, W., Terhalle, W. and Dress, A. A Quantitative Analysis of Basic Helix-Loop-Helix Domain Containing Proteins. In preparation.

Brack-Werner, R., Görblich, T., Werner, T., Chiodi, F., Gürtler, L., Eberle, J., Schön, A., and Erfle V. (1994). Sequence analysis of neuroinvasive and blood-derived HIV-1 nef genes, Technical Advances in AIDS Research in the Human Nervous System (Eds. E. O. Major, J. A., Levy), Plenum Press, New York, 189-204.

Brigelius-Flohé, R., Aumann, K.-D., Blöcker, H., Gross, G., Kiess, M., Klöppel, K.-D., Maiorino, M., Roveri, A., Schuckelt, R., Ursini, F., Wingender, E. and Flohé, L. (1994). Phospholipid hydroperoxide glutathione peroxidase: genomic DNA, cDNA, and deduced amino acid sequence. J. Biol. Chem. 269, 7342-7348.

Fabian, H., Hölzer, W., Heinemann, U., Sklenar, H. and Welfle, H. (1993). Conformation of $d(GGGATCCC)_2$ in crystals and in solution studied by X-ray diffraction, Raman spectroscopy and molecular modeling. Nucleic Acids Res. 21, 569-576.

Graw, J., Liebstein, A., Pietrowski, D., Schmitt-John, T. and Werner, T. (1993). Genomic sequences of murine gB- and gC-crystallin genes: promoter analysis and complete evolutionary pattern of mouse, rat, and human g-crystallins. Gene 136, 145-156.

Haag, F. A., Kuhlenbäumer, G., Koch-Nolte, F., Wingender, E. and Thiele, H.-G. (1996). Structure of the gene encoding the rat T cell ecto-ADP-ribosyltransferase RT6. J. Immunol. 157, 2022-2030.

Kütemeier, G., Höhne, W., Werner, T., Schuh, R. and Mocikat, R. (1994). Assembly of humanized antibody genes from synthetic oligonucleotides using a single-round PCR. Biotechniques 17, 242-246.

Mautner, J., Joos, S., Werner, T., Eick, D., Bornkamm, G., W. and Polack, A. (1995). Identification of two enhancer elements in the 3' region of the human c-myc gene. Nucleic Acids Res. 23, 72-80.

Möritz, A., Grzeschik, K.-H., Wingender, E. and Fink, E. (1993). Organization and sequence of the gene encoding the human acrosin-trypsin inhibitor (HUSI-II). Gene 123, 277-281.

Pietrowski, D., Durante, M., J., Liebstein, A., Schmitt-John, T., Werner, T. and Graw, J. (1994). Lens a-crystallin interacts specifically with murine gD/E/F-crystallin promoter. Gene 144, 171-178, 1994.

Ramser, J., Weising, K., Lopez-Peralta, C. , Terhalle, W., Terauchi, R. and Kahl, G. Molecular marker-based taxonomy and phylogeny of Guinea Yam (Dioscorea rotundata-D. cayenensis). Subbmitted to Genome.

Sedlmeier, R., Werner, T., Kieser, H., M., Hopwood, D., A. and Schmieger, H. (1994). tRNA genes of Streptomyces lividans: new sequences and comparison of structure and organization with those of other bacteria. J. Bacteriology 176, 5550-5553.

Weitere, im Text zitierte Publikationen:

Frech, K., Herrmann, K. and Werner, T. (1993). Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. Nucleic Acids Res. 21, 1655-1664.

Lavery, R., Sklenar, H., Zakrzewska, K. and Pullman, B. (1986). The flexibility of nucleic acids: (II) the calculation of internal energy and applications to mononucleotid repeat DNA. J. Biomol. & Dyn. 3, 989-1014.

Lavery, R. and Sklenar, H. (1988). Calculation of helicoidal parameters for irregular nucleic acid structures: the CURVES algorithm. J. Biomol. Struct. & Dyn. 6, 63-91.

Lavery, R. and Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. J. Biomol. Struct. & Dyn. 6, 655-667.

Wingender, E. (1988). Compilation of transcription regulating proteins. Nucleic Acids Res. 16, 1879-1902.
Wingender, E. (1993). Gene Regulation in Eukaryotes. VCH Weinheim.