5. Diskussion

5.1. Die Klon-Datenbank *CloneBase*

Zur Durchführung von Hochdurchsatz-Experimenten wie matrixCGH und *Expression Profiling* über Microarrays werden Sammlungen aus vielen tausend Einzelfragmenten benötigt. Für die Archivierung, Annotation und Informationsbereitstellung einer solchen Sammlung wurde das *CloneBase*-System entwickelt.
Die Datenbanksysteme *CloneX* and *CloneY* wurden erstmals Mitte 2000 vorgestellt. Seitdem wurden regelmäßig Verbesserungen in Bezug auf Stabilität, Abfragemöglichkeiten und Annotationsdaten vorgenommen. Als Systeme mit vergleichbarer Zielstellung können die Datenbank des *IMAGE*-Konsortiums, sowie das *FANTOM*-Projekt (The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team, 2002) genannt werden. Beide verwalten umfangreiche Sammlungen von cDNA-Klonen, die sich per Internet abfragen lassen. Die Systeme sind jedoch nicht für eigene Projekte verfügbar. Auch um die speziellen Informationen sowohl genomischer als auch exprimierter Fragmente in einer Datenbank verwalten zu können, war es notwendig, eine eigene Lösung zu entwickeln, die nach Bedarf auch problemlos weiterentwickelt werden kann. Das *CloneBase*-System ist auf die Bedingungen zugeschnitten, eine heterogene Sammlung von Klonen zu verwalten, automatisiert zu annotieren und einer verteilten Benutzergruppe zugänglich zu machen. Das System stellt einerseits die Datenbasis für das Labordatensystem *QuickLIMS* zur Produktion von Microarrays dar, indem es die Identifikationsnummern verfügbarer Klone und deren Verteilung auf Mikrotiterplatten speichert. Anderseits liefert es die Ausgangsinformationen für die funktionelle Interpretation der Endergebnisse, die aus Experimenten mit diesen Klonen gewonnen werden, durch *FACT*.

Eingesetzte Technologien

Die Daten der vorhandenen Klonsammlungen sind die Grundlage diverser Forschungsprojekte von hoher klinischer und biologischer Relevanz. Das System ist daher auf einem leistungsfähigen spezialisierten Servercomputer installiert, der als Betriebssystem *Linux* besitzt (Sicherheits- und Leistungsaspekt). Sämtliche eingesetzten Software-Anwendungen (*mySQL*, *Apache*, *Java*-

Entwicklungsumgebung, *Perl*- und *PHP*-Interpreter) sind frei erhältlich und in Bezug auf Geschwindigkeit und Konfiguration gut geeignet. Der *mySQL-Server* wird beispielsweise auch vom *Ensembl*-Projekt benutzt, um über 80 GB Genomdaten zu verwalten, die weltweit abgerufen werden. *Apache* wird von über 70% aller Internet-Server verwendet (http://www.PRNewswire.com). Durch den Einsatz einer Skriptsprache (PHP) erfolgt die Verarbeitung der programmierten Funktionalität bereits auf dem Servercomputer. Sämtliche Daten und andere Programme können daher auf dem Server genutzt werden und erst nach Abschluss von Berechnungen und Verarbeitungen werden die benötigten Daten über das Netzwerk gesendet. Hierdurch verringert sich die so genannte Response-Zeit (Zeitintervall vom Beginn einer Anfrage bis zur Rückgabe eines Ergebnisses) und die Netzwerklast wird verringert (Datenmenge, die über das Intranet/Internet übertragen wird). Die Ausführungsgeschwindigkeit der Serveranwendung kann durch den Einsatz eines leistungsfähigeren Computers bei Bedarf ohne weitere Veränderungen der Anwendung gesteigert werden.

Datenbank- und Programm-Design

Beim Datenbank-Design ist im Allgemeinen auf eine hohe Effizienz in Bezug auf Speichernutzung und Abfragegeschwindigkeit zu achten. Das relationale Design der *CloneBase*-Datenbank bewirkt neben erhöhter Effizienz auch, dass sich die Geninformationen unabhängig für alle Klone aktualisieren lassen.

Die Internet-Abfrageseiten wurden in Kollaboration mit Anwendern innerhalb der Projekte so weiterentwickelt und gestaltet, dass ein höchstmöglicher Bedienerkomfort geschaffen werden konnte, ohne Einbußen an Funktionalität akzeptieren zu müssen. Die Abfragemöglichkeit durch das Hochladen einer Liste (z.B. mit Klonnamen) ist ein Beispiel hierfür, das direkte Generieren einer Datei zum Abspeichern nach einer umfangreichen Suche ein anderes. Das rasche Auffinden der gesuchten Informationen wird hierdurch wesentlich verbessert.

Datenbasis und Aktualisierung

Zu Beginn des Projektes wurde die Datenbank *euGene* als Datenbasis eingesetzt. Da diese Datenbank als *Meta-Datenbank* Daten von unterschiedlichen *Primärdatenbanken* vereinigt, kann eine umfassende Annotation von einer einzelnen Datenquelle erfolgen. Im Verlauf der letzten Jahre entstand und etablierte sich das

Genom-Datenbanksystem *Ensembl* als umfangreiche öffentlich verfügbare Annotationsquelle. Die Annotation der Klon- und Gendaten wurde von *euGene* auf *Ensembl* umgestellt, da einerseits eine ausgereifte Datenbank-Schnittstelle (API) zu *Ensembl* in den Sprachen Perl und Java existiert und andererseits mit dem öffentlichen Server vom Wellcome Trust Sanger Centre / EBI (ensembldb.ensembl.org) ein leistungsfähiges System per (Datenbank-) Internetverbindung zur Verfügung steht. Das Beziehen, Parsen und lokale Nachbilden des Datenbestandes entfallen dadurch. Die Ausführung einer Annotation ist in monatlichem Rhythmus sinnvoll, da *Ensembl* monatlich zumindest in Teilen aktualisiert wird. Die Daten lassen sich gezielt einzeln oder in Folge abrufen.
Um eine möglichst vollständige Annotation erreichen zu können, wird im Script sowohl mit Mitteln der leistungsfähigen Perl-API und per direktem SQL-Befehl in der *Ensembl*-Datenbank, als auch in der *Golden-Path* Datenbank der *University of California* nach Informationen zu den Klonen gesucht. Diese werden dann durch Genannotationen aus *Ensembl* erweitert. Verifizierbar sind die Annotationsdaten prinzipiell nur mit großem Aufwand, da *Ensembl* jedoch eine ganze Reihe unterschiedlicher Annotations-Algorithmen einsetzt (Curwen *et al.*, 2004) und Programmfehler von Nutzern der industriellen und akademischen Forschung rasch entdeckt und gemeldet werden (vergl. Mailingliste der Programmierer *ensdev.emsebl.org*), kann die Datenbasis als gesichert betrachtet werden. An Daten und Datenbank-Software von *Ensembl* wird aktiv weitergearbeitet und Erweiterungen sind schnell in die *CloneBase* integrierbar. Die Annotationsinformationen können mit dem vorgestellten System vollautomatisch auf dem aktuellsten Stand gehalten werden.

Anwendungen in der biologisch-medizinischen Forschung

Die Daten der Klonsammlung werden täglich von Mitarbeitern innerhalb und - über eine separate Abfragemaske im Internet - auch außerhalb der Arbeitsgruppe abgerufen und zur Auswertung von Microarray-Experimenten sowie bei der Erstellung neuer Arrays genutzt. Die *CloneBase* unterstützte dadurch bereits maßgeblich eine Reihe verschiedener Forschungsvorhaben (Schlingemann *et al.*, 2003, Neben *et al.*, 2003, Neben *et al.*, 2004, Korshunov *et al.*, 2004, Wrobel *et al.*, eingereicht, Hummerich *et al.*, eingereicht). Sie stellt ferner weiterführenden Anwendungen für die Microarrayproduktion (*QuickLIMS*) und Auswertung (*FACT*) eine strukturierte und effizient abfragbare Datenbasis zur Verfügung.

5.2. Die Prozess-Datenbank *Quick-LIMS*

Die Produktion von Microarrays mit mehreren tausend Zielstellen erzwingt den Einsatz von maschineller Unterstützung (Benton 1996). Der eingesetzte Pipettier-Roboter *MiniTrak* kann bei optimaler Auslastung einen Durchsatz von ca. 12000 Proben pro Woche erreichen (Wrobel 2004). Sowohl die koordinierte Steuerung der Maschine als auch das hohe Datenaufkommen lassen sich ohne ein spezialisiertes Labordatensystem nicht sinnvoll bewältigen.
Ein spezialisiertes Microarray-LIMS mit der erwähnten Funktionalität wurde im akademischen Bereich bisher nicht vorgestellt. Kommerzielle Lösungen sind kostenintensiv und müssen für die Integration in das System grundlegend angepasst werden (z.B. *Partisan* von Clondiag, Jena oder *Clonetracker* von Biodiscovery, El Segundo/U.S.A.). Veröffentlichte Datenbanksysteme für Microarrays haben entweder die bei der Produktion entstehenden Daten nicht oder nur teilweise integriert (Brazma *et al.*, 2003, Gollub *et al.*, 2003) oder erlauben nicht die gewünschte Interaktion mit einem Laborroboter (Fellenberg *et al.*, 2002).
Es wurde daher ein LIMS programmiert, welches sämtliche anfallende Daten verwalten kann, und sowohl Roboter als auch Benutzer durch den Herstellungsprozess der Microarrays leitet. Die flexible Definition des Prozessablaufs und die enge Einbindung des Roboters stellen Neuerungen in diesem Gebiet dar. Das System kann daher als Prototyp eines Labordatensystems dienen, wie es in unterschiedlichen wissenschaftlichen Umgebungen genutzt werden kann.

Eingesetzte Technologien

Das *QuickLIMS*-System wurde aus unterschiedlichen Gründen als Microsoft Access™ Datenbank-Programm entwickelt. Einer dieser Gründe waren die technischen Grundvoraussetzungen. Access-Anwendungen verwenden die Programmiersprache *Visual Basis for Applications* (*VBA*), welche auch von den Steuerungsfunktionen des Pipetier-Roboters genutzt wird. Auch der relativ unkomplizierte Aufbau von VBA war ein Grund, dieses Programm zu verwenden. Dadurch können zu einem späteren Zeitpunkt nötige Veränderungen im Programmablauf auch von versierten Anwendern ohne lange Einarbeitungszeit umgesetzt werden. Des Weiteren hatte die Produktion von Microarrays bereits begonnen und mit den Möglichkeiten, die Access bietet, kann relativ schnell eine funktionsfähige Datenbank mit ersten Benutzermasken erstellt

werden.

Die Verwendung von Microsoft Access bringt jedoch auch Nachteile mit sich. Bei einem hohen Datenaufkommen verlängert sich die Response-Zeit des Systems stark. Mit einem derzeitigen Datenbestand von 3700 Platten und 250000 Klon-Informationen verlangsamt sich der Bearbeitungsprozess für den Benutzer merklich. Es muss daher in regelmäßigen Abständen, bzw. bei Bedarf ein Teil der Daten ausgelagert werden. Ein weiterer Nachteil stellt die Inkompatibilität von Access gegenüber anderen Betriebssystemen als Microsoft Windows dar. Um die Daten für andere Anwendungen, wie z.B. das Analysesystem *FACT*, problemlos zur Verfügung zu haben, wird täglich mittels eines Perl-Skripts der gesamte Datenbestand in eine *mySQL*-Datenbank auf dem (Linux-)Server übertragen.

Datenbank- und Programm-Design

Das Prozess-System ist *Platten-orientiert*, d.h. zentrale Informationseinheit sind die virtuellen Repräsentationen der Mikrotiterplatten. Ein sechsstelliger Barcode identifiziert jede Platte im System als *Master-Plate*, *Process-Plate* oder *Spotting-Plate*. Diese Aufteilung erlaubt es, von einer Platte (*Master-Plate*) mehrere Replikate als Prozessplatten zu erzeugen, ohne die Information über die gemeinsame Herkunft zu verlieren, und ohne Daten über die Plattenbelegung neu eintragen zu müssen. Die ursprüngliche Information über Platten und Klone wird der Klondatenbank *CloneBase* entnommen.

Des Weiteren ist das LIMS *Protokoll-orientiert*, d.h. die Schrittfolge des Programms richtet sich nach dem definierten Ablauf. Dieser ist in einer eigenen Tabelle der Datenbank gespeichert, also nicht im Programmcode fixiert. Veränderungen der Parameter und Abläufe können daher schnell und problemlos umgesetzt werden, auch die Datenformate sind in der Tabelle als „Meta-Informationen“ abgelegt.

Durch diese Struktur und durch einen dieser Flexibilität angepassten Programmcode ist es möglich, Formulare dynamisch dem jeweiligen Protokollschritt entsprechend zu generieren. Für keinen der Prozessschritte muss daher eine eigene Datenmaske gespeichert werden, was die Komplexität verringert und eine hohe Flexibilität herbeiführt.

Das System erreicht eine lückenlose Speicherung der Daten und korrekte Weiterbehandlung der Mikrotiterplatten. Sämtliche Daten des Herstellungsprozesses können in die spätere Auswertung der Experimente mit einbezogen werden und

ermöglichen die Qualitätssicherung und Fehlerrückverfolgung. Es wurde eine hohe Prozesssicherheit erreicht, mit der auch ein ungeübter Nutzer durch das Protokoll geführt werden kann.

Anwendungen in der biologisch-medizinischen Forschung

QuickLims wurde innerhalb der Abteilung zur Produktion von Microarrays eingesetzt, insbesondere zur Untersuchung der Expressionsprofile von Meningiomen (Wrobel *et al.*, eingereicht) und von *Non-Melanom* Hautkrebs (Hummerich *et al.*, eingereicht). Produziert wurden dabei Arrays mit ca. 12000 humanen Fragmenten, sowie murine Arrays mit 40000 (LION Klone), bzw. 30000 (NIA Klone) Sequenzen.

5.3. Das Primergenerierungs-Werkzeug *AutoPrime*

Die Verifikation der erzielten Microarray-Ergebnisse durch die RQ-PCR wurde mit der Anwendung *AutoPrime* zu einem weiteren Teil automatisiert. Es existiert zwar eine Vielzahl von Programmen zum Design von Primern (z.B. Haas *et al.*, 1998, Thareau V *et al.*, 2003, Rozen und Skaltsky, 2000), inzwischen sogar spezielle Datenbanken für Primersequenzen (Pattyn *et al.*, 2003, Wang *et al.*, 2003). Die Ergebnisse der Genomsequenzierung sind ebenfalls in öffentlich Datenbanken zugänglich (z.B. Pruitt *et al.*, 2000). Die Kombination beider Systeme fehlte jedoch bisher völlig, die Sequenzen mussten manuell von den Datenbank-Internetseiten an die Primerdesign-Software übergeben werden. Diese Lücke wird durch *AutoPrime* für RQ-PCR-Primer geschlossen, indem mit Hilfe von Perl-Skripten in der *Ensembl*-Datenbank nach den notwendigen Sequenzinformationen gesucht wird und diese direkt an das Primerdesign-Programm *Primer3* übergeben werden. Über die gezielte Platzierung der Primer auf Exon-Grenzen oder in unterschiedliche Exone und über die Nutzung der Information über genomische Wiederholungssequenzen (*Repeat-Library*) kann *AutoPrime* die Kontamination durch genomische PCR-Produkte verringern. Nicht vermeiden lassen sich natürlich weiterhin Kreuzhybridisierungen mit homologen Sequenzen. Auch das Vorhanden sein von Pseudogenen kann evtl. zu Problemen bei der automatisierten Verarbeitung führen. Dies müsste in einem größeren Ansatz getestet werden.

Im Vergleich zur herkömmlichen Methode wird durch *AutoPrime* die manuelle

Interaktion des Benutzers auf ein Minimum begrenzt. Dies erleichtert ihm einerseits die Arbeit und verringert andererseits die Fehlerrate.

Nach dem Fertigstellen der Anwendung *AutoPrime* wurde ein ähnliches System von O. J. Marshall vorgestellt (*PerlPrimer*, Marshall, 2004). *PerlPrimer* hat insgesamt eine Reihe Funktionen, die über die Aufgabenstellung von *AutoPrime* hinausgehen. Es können beispielsweise unterschiedliche PCR-Methoden ausgewählt werden und Sequenzen können direkt zum *BLAST*-Programm des NCBI (http://www.ncbi.nlm.nih.gov/BLAST) gesendet werden. Es ist jedoch nicht per Web-Oberfläche nutzbar.

Als Alternative können Gen-spezifische Primer für die RQ-PCR kommerziell bestellt werden. Die Sequenzen sind dabei vorberechnet und in Datenbanken gespeichert (*Essay-On-Demand*, Applied Biosystems, Darmstadt).

Eingesetzte Technologien

Die Anwendung wurde durchgehend in der Sprache Perl entwickelt. Dadurch kann in der gleichen Umgebung einerseits die dynamische Erzeugung der Internet-Oberfläche durchgeführt werden und andererseits mit direkter Anbindung an Ensembl (*Ensembl* Perl-API) und mit einem Systembefehl zum Start von *Primer3* das *AutoPrime*-Hauptprogramm gesteuert werden. Über die Web-Oberfläche kann *AutoPrime* ohne Anforderungen an das Betriebssystem, an installierte Anwendungen und an bestimmte Rechenleistungen ausgeführt werden. Da der zum Teil zeitintensive Schritt des Heraussuchens und Aufbereitens der Ausgangssequenzen bereits von der Benutzeroberfläche (der Eingabe-Maske) abgekoppelt wird (*forking*), kann die Programmausführung auf dem Server ablaufen, ohne den Computer des Benutzers aufgrund hoher Auslastung zu blockieren, bzw. ohne das die Gefahr besteht, das der Internetserver wegen Überschreitung eines Zeitlimits die Bearbeitung abbricht. Mittels einer Protokolldatei kann der Fortschritt des Hauptprogramms jedoch weiterhin auf der Internetseite des Nutzers gezeigt werden. Das Hauptprogramm ist zudem technisch von der Internet-Anbindung separiert und kann dadurch auch per Befehlszeile oder über andere Programme genutzt werden. Letzteres wird darüber hinaus durch die primäre Ausgabe des Ergebnisses als XML-Datei erleichtert.

Anwendungen in der biologisch-medizinischen Forschung

AutoPrime wurde unter der Adresse http://www.AutoPrime.de vorgestellt (Wrobel *et*

al., 2004) und wird seitdem von unterschiedlichen Wissenschaftlern genutzt. Innerhalb der Arbeitsgruppe wurde es bisher in zwei Projekten erfolgreich angewendet (Wrobel *et al.*, eingereicht und Hummerich *et al.*, eingereicht).

5.4. Funktionelle Analyse von Experimenten (*FACT*)

Die Anwendung von hochparallelen Methoden in der molekularbiologischen Forschung erfordert neue Ansätze zur Aufarbeitung der großen Datenmengen. Die Integration neuer Information aus unterschiedlichen Forschungsbereichen ist dabei eine Grundvoraussetzung. Nur durch die Automatisierung einer flexiblen Analyse, wie sie *FACT* durchführt, wird es möglich, biologische Zusammenhänge zu erkennen und z.B. für medizinische Aussagen zu nutzen. Hochdurchsatz-Methoden wie DNA-Microarray-Analysen geben dabei die Möglichkeit, genomweite Untersuchungen zu realisieren und mit diesem *Screening* weitere Experimente zielgerichtet auszuführen. Um diese Zielrichtung besser einschätzen zu können, liefert *FACT* die Möglichkeit, in einer explorativen Datenanalyse interessante Aspekte in den Datensätzen schneller zu identifizieren.

Integration von Datenquellen

Die Integration möglichst umfangreicher Annotationsdaten spielt eine Schlüsselrolle bei der Auswertung von Hochdurchsatzexperimenten. Spätestens seit dem Abschluss der Sequenzierung der Genome von Maus und Mensch und mit stetig anwachsenden Kenntnissen über die damit verbundenen biologischen Zusammenhänge, wächst die Auswahl verfügbarer Daten mit hoher Geschwindigkeit. Das Analysesystem *FACT* wurde für die Interpretation von experimentellen Ergebnissen mit Hilfe von Annotationsdaten und die Zusammenführung von relevanten Informationen aus heterogenen Datenquellen entwickelt. Es existieren bereits unterschiedliche Ansätze zur Integration von heterogenen Daten. Das bekannteste System ist SRS (*Sequence Retrieval System*, Etzold *et al.*, 1996), ein auf indizierten Textdateien basierendes System, welches zur parallelen Abfrage von vielen Datenbanken genutzt wird. Wie auch bei *FACT*, werden die Datenquellen über individuelle Parser angesprochen, deren Integration jedoch relativ komplex ist (Leser und Rieger, 2003). Bei *DiscoveryLink* (Haas *et al.*, 2001) erfolgt die Integration als eigene (virtuelle) Tabellen,

die jeweils über „Wrapper“ angesprochen werden. Als Wrapper bezeichnet man eine Funktion, die eine Kommunikation zwischen unterschiedlichen Datenmodellen ermöglicht, indem sie beim Zugriff eine Transformation der Schemata vornimmt. Beide Konzepte ermöglichen die Einbindung von heterogenen Datenquellen und haben sich in der Praxis bewiesen. Die Funktionsweise von *FACT* geht über diese Aufgaben hinaus. Auch experimentelle Daten unterschiedlicher Quellen können hiermit integriert werden. Es sind ferner Analysemodule eingebunden, welche sämtliche Daten direkt miteinander in Verbindung bringen können. Die Integration unterschiedlicher Quellen als experimentelle Ausgangsdaten oder als Annotationsdaten erfolgt bei *FACT* durch die Erstellung eigener Module, die nach einem vorgegebenen Prototyp entwickelt werden können und deren Struktur relativ simpel sein kann. Sie können im laufenden Betrieb von jedem Benutzer (mit entsprechenden Rechten) auch über das Netzwerk auf den Server geladen und direkt verwenden werden. Als „Datenquell-Adaptoren“ führen sie die Transformation von speziellen Quellen, Formaten und Inhalten zu einem gemeinsamen Datenkonzept, das in *FACT* genutzt wird, durch. Dies steht in Analogie zu den Wrappern in *DiscoveryLink*, jedoch erfolgt bei *FACT* eine Speicherung der Daten im gemeinsamen Schema, sodass sie allen weiteren Annotations- und Analysemethoden in gleicher Art und Weise zur Verfügung stehen. Diese Generalisierung der unterschiedlichen Daten ist ein Kernkonzept, welches das System auszeichnet.

Integration von Analysemethoden

Auch neue Analysemethoden werden in diesem Stil als Module eingebunden. Da derzeit viele neue Einzelansätze zur Datenanalyse entwickelt werden, ist die Möglichkeit, unterschiedliche Methoden einbinden und mit den gleichen Daten testen zu können, sinnvoll. *GoMiner* (Zeeberg *et al.*, 2003) ist eine Anwendung, zur Erweitern von Genlisten mit Annotationen der GeneOntology. *OntoExpress* (Khatri *et al.*, 2002) und *GeneMerge* (Castillo-Davis *et al.*, 2003) gehen einen Schritt weiter und suchen signifikant erhöhte Vorkommen von GO-Annotationen. Sie nutzen dabei die hypergeometrische Verteilungsfunktion. *EASE* (Hosack *et al.,* 2003) verwendet dazu den Fisher’s Exakt-Test und kann auch weitere Annotationen untersuchen (z.B. die chromosomale Lokalisierung).

Das *Flexible Annotation and Correlation Tool* ermöglicht die Integration dieser Algorithmen und beliebiger weiterer Funktionalitäten. Es ist bereits möglich, GO-

Begriffe von unterschiedlichen Quellen zu benutzen, z.B. um eine maximale Annotationen zu erhalten oder um die Unterschiede zu sehen. Eine Detektion signifikanter Anhäufungen können in *FACT* über die hypergeometrische Verteilungsfunktion auf sämtliche Datentypen angewendet werden.
Um neue Erkenntnisse aus den experimentellen Daten gewinnen zu können ist es sinnvoll, verschiedene Annotationen und Analysemethoden in Kombination anzuwenden. Bisher unbekannte Zusammenhänge, wie z.B. Signalketten in der Zelle, sind nicht mit Standardmethoden erkennbar. Ergebnisse unterschiedlicher Herkunft benötigen verschiedene Analyse-Ansätze, um eine neue Sichtweise auf die Daten zu erhalten. Die Arbeit mit *FACT* kann z.B. sowohl mit Gen-Symbolen, Klon-IDs als auch mit Affymetrix-IDs begonnen werden, der Annotations- und Analyseablauf ist nicht festgelegt. Derzeit geläufige Programme erwarten dagegen eine Festlegung auf bestimmte Datentypen und bieten nur bestimmte Methoden an.

Offenheit für neue Entwicklungen

Das *FACT*-System besitzt eine Struktur, die das flexible Speichern von heterogenen Informationen erlaubt. Es können eigene Datenquellen definiert und deren Bedeutung aus einer Datentyp-Definitions-Tabelle gewählt werden. Vor dem Hintergrund stetig wachsender Informationsquellen und ganz unterschiedlicher Datenformate ist dies eine Grundvoraussetzung für ein zukunftsorientiertes System. Daher wird in *FACT* auch der modulare Aufbau stark betont. Spezialisierte Funktionen werden abgekapselt und liefern als Ausgabe standardisierte Formate. Es stellt ein System zur Verfügung, in welchem heterogene Daten durch Modularisierung und Abstraktion zusammengeführt werden können. Unter anderem fließen die in der *CloneBase*- und der *QuickLIMS*-Datenbank gespeicherten Informationen mit in die Analyse ein.
Außer den abteilungsinternen Ressourcen (Klon- und CGH-Datenbank) sind sämtliche Komponenten frei erhältlich. Neben der Hauptinstallation am Deutschen Krebsforschungszentrum (http://www.factweb.de), welche sämtliche Funktionalitäten mittels der Internet-Oberfläche zur Verfügung stellt, kann das System daher problemlos auch in anderen Institutionen installiert werden. Dieses *Open-Source*-Konzept ist entscheidend, um eine effiziente Weiterentwicklung zu gewährleisten. Durch die Gestaltung von spezifischen Modulen von einzelnen Benutzern kann so ein System evolvieren und Bestand haben.

Eingesetzte Technologien

Das Hauptprogramm, die Mehrheit der Module und auch die Web-Anwendung von *FACT* sind in der Scriptsprache *Perl* geschrieben. Perl ist eine vor allem im Bereich der Bioinformatik sehr weit verbreiteten Programmiersprache, die sich gut zur Interaktion im Internet und mit Textdateien eignet und dadurch ideal zum Einlesen der Annotationsdaten verwendet werden kann. Es gibt bereits zahlreiche Programm-Module in der Sprache Perl, die in die Entwicklung mit eingebunden und genutzt werden können (z.B. *BioPerl*: Stajich *et al.*, 2003; *DBI*: http://cpan.org; *Ensembl*-Module: Stabenau *et al.*, 2004). Dieser Aspekt ist für *FACT* entscheidend, da nur die Nutzung und problemlose Integration von existierenden Algorithmen und Funktionalitäten für ein Projekt dieser Größe Erfolg versprechend ist.

Anwendungen in der biologisch-medizinischen Forschung

Das *Flexible Annotation and Correlation Tool* ist eine relative junge Entwicklung, wurde jedoch bereits für verschiedene Fragestellungen erfolgreich angewendet. Die Funktion zur Korrelation zwischen Expressions- und CGH-Daten wurde im Rahmen eines Projektes zur Untersuchung der Entwicklung von Meningiomen (Wrobel *et al.*, eingereicht) entwickelt. Die Darstellung von Daten im genomischen Kontext wurde zur Datenanalyse der Untersuchung von Medulloblastomen mittels matrixCGH (Mendrzyk *et al.*, eingereicht) angewendet. Für die Evaluierung der Daten zur Progression von Basalzell Karzinomen und Squamouszell Karzinomen am Modell der chemisch induzierten Karzinogenese an der Maus (Hummerich *et al.*, eingereicht) wurden umfangreiche Annotations- und Analysefunktionen von *FACT* genutzt (siehe Kapitel 4.4.2.). Diese lieferten entscheidende Hinweise für die Interpretation des Experimentes (siehe Kapitel 5.5.).

5.5. Untersuchung der Pathomechanismen von *Non-Melanom* Hautkrebs

Mit Hilfe des beschriebenen Systems konnten neben der Untersuchung von Hirntumoren (Glioblastome, Medulloblastome) und hämatologischen Fragestellungen (Zelllinie HL60, Akute Myeloische Leukämie), eine umfassende Studie über die Pathomechanismen der Entstehung von Papillomen und der Progression zu

Karzinomen der Haut durchgeführt werden. Diese umfangreiche Studie wurde mit zwei murinen 15000 bzw. 20000 Fragment-Microarrays durchgeführt, welche mit Hilfe des *CloneBase*- und *QuickLIMS*-Systems entstanden. Bei der anschließenden Analyse der Daten wurden durch Funktionen von *FACT* die umfassende Annotation der Kandidatengene erreicht und eine Interpretation der Ergebnisse ermöglicht. Am Modell der Haut der Maus konnten Gene und Genfamilien als Kandidaten für diese Entwicklung identifiziert werden. Es handelt sich dabei um Gene mit Zellwachstums- und Zellteilungsfunktionen, sowie Mitglieder der *S100*-Genfamilie. Von den 21 humanen *S100*-Proteinen, die sich durch ein Kalzium-bindendes EF-Sequenzmotif charakterisieren, sind 14 im Epidermalen Differenzierungskomplex (*EDC*) in der chromosomalen Region 1q21.3 lokalisiert (Abb. 23). Gene des EDC besitzen eine Mediatorfunktion innerhalb von zellulären Signalwegen, werden von extrazellulären Stimuli aktiviert und sind für die Ausdifferenzierung der Keratinozyten in der Epidermis entscheidend. Ihre Rolle bei der Pathogenese epidermialer Krankheiten beinhaltet unter anderem Psoriasis, Wundheilung, Hautkrebs und Entzündungsreaktionen (Eckert *et al.*, 2004).

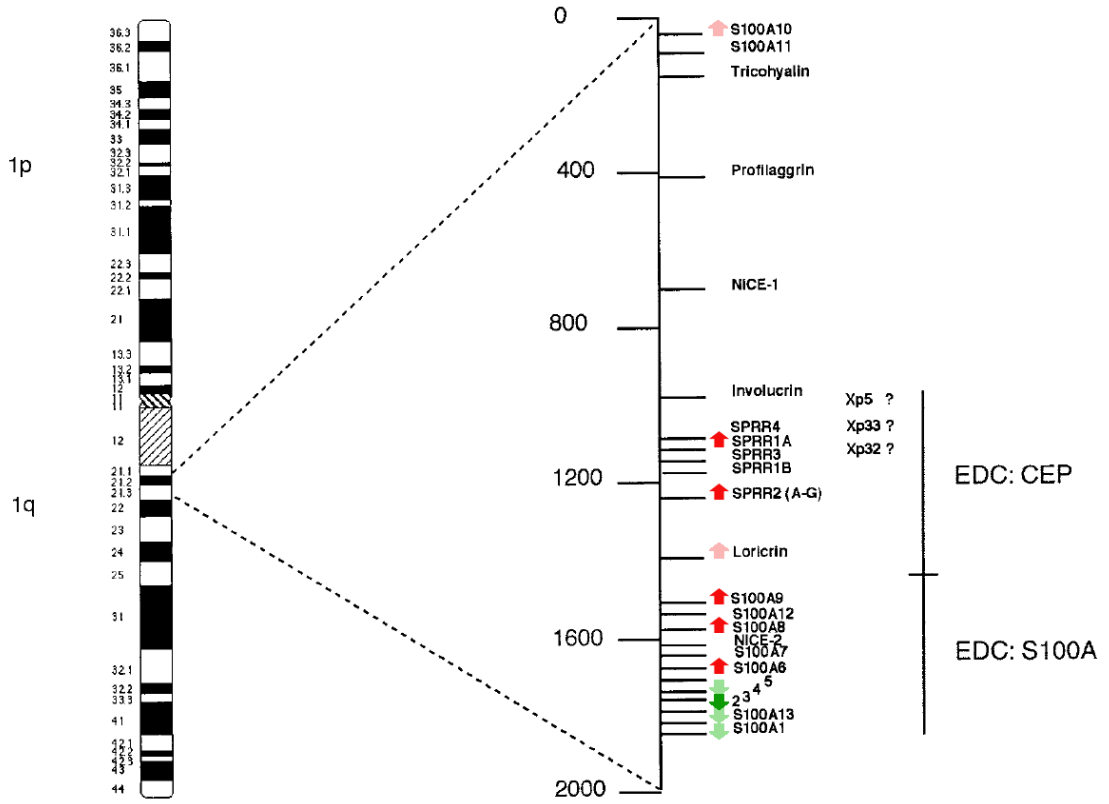

Abb. 23: Karte der 2 MB *EDC*-Region in der chromosomalen Bande 1q21

Gezeigt sind alle bekannten Gene sowie die differnezielle Expression der Gene, die auf den Microarrays repräsentiert waren. Logarithmierte Ratios: dunkelrot > 1.0, hellrot > 0, hellgrün < 0, dunkelgrün < -1.

Eine Überexpression der *S100*-Gene bei der Entstehung von *Non-Melanom* Hautkrebs ist daher von großer Relevanz, der Lokus 1q21 stellt eine Kandidatenregion dar. Für die Verifikation der Expressionswerte mittels RQ-PCR wurde *AutoPrime* zur Generierung der Primersequenzen genutzt.

5.6. Das Gesamtsystem

Die im Rahmen dieser Arbeit entwickelten Programme sind trotz unterschiedlicher Einzelfragestellungen und verschiedener eingesetzter Technologien eng miteinander verbunden. Das Annotations- und Analsyse-System *FACT* ermöglicht die Integration von Datenquellen und Funktionen in einer konzeptionell neuen Flexibilität. Indem es nicht nur Daten aus externen Quellen, sondern auch der Klondatenbank *CloneBase* und des Prozesssystems *QuickLIMS* aufnehmen und einer Analyse zuführen kann, bauen die vorgestellten Systeme aufeinander auf. *QuickLIMS* greift seinerseits bei der Erstellung neuer Klon-Platten für die Microarray-Produktion auf Informationen zu, die in der Klondatenbank gespeichert sind. Mit weiteren in der Abteilung entwickelten Softwarelösungen (Microarray-Ergebnisdatenbank und Analyseskripte) wurde somit ein System geschaffen, welches als Datenbank- und Analyse-Netzwerk den experimentellen Ablauf begleitet. Es ermöglichte die Bearbeitung unterschiedlicher Fragestellungen zur Untersuchung der Entstehung und Progression von Tumor-Entitäten.

Der Einsatz der automatisierten und computergestützten Methoden bedeutet neben einer vereinfachten Handhabung auch die Reduktion potentieller Fehler im Ablauf und in der Auswertung. Dies ist vor allem in der klinisch-relevanten Forschung von größter Bedeutung. Viele experimentelle und analytische Ansätze werden durch das enge Zusammenwirken von Biologie und Informatik überhaupt erst ermöglicht. Die Klon-Datenbank stellt als Archivierungs- und Suchwerkzeug die Grundinformationen in einer verlässlichen Form bereit. Das Labordatensystem leitet Roboter und Benutzer sicher durch den Herstellungsprozess von Microarrays. Es speichert dabei alle relevanten Daten. Mit *AutoPrime* wird auch der Prozess der Verifikation der erzielten Microarray-Ergebnisse mit der Methode der RQ-PCR zu einem weiteren Teil automatisiert. *FACT* ermöglicht schließlich den Zugang zu Datenquantitäten, wie sie der biologischen Forschung heute zur Verfügung stehen. Das Zusammenführen

heterogener Datensätze und die explorative Anwendung neuer Analysemethoden ist ein Schlüsseltechnik, die zu einer qualitativ besseren Analyse (mehr Daten und aktuellere Daten), zu verkürzten Entwicklungszeiten (Nutzung von existierenden Informationen) und zur neuen Erkenntnissen (Korrelationen zwischen zuvor unverbundenen Datensätzen) führen kann.