Problemseminar Bio-Datenbanken
WS 2002/2003

Pathway-Datenbanken

Bearbeiter: Jörg Lehmann

Betreuer: Sergej Melnik

Prof. Dr. Erhard Rahm

Universität Leipzig
Institut für Informatik

Januar 2003

Inhaltsverzeichnis

1. Einleitung

Während der vergangenen Jahrzehnte wurden in den Biowissenschaften entscheidende Fortschritte erzielt. Von besonderer Bedeutung ist dabei vor allem die Sequenzierung des Genoms verschiedener Organismen, insbesondere das des Menschen, die die notwendige Grundlage zur gezielten Analyse der Erbanlagen bietet. Dazu kam bisher vor allem der reduktionistische Analyseansatz zur Anwendung: Das biologische Gesamtsystem wird in feinere Teilsysteme zerlegt, deren einzelne Bestandteile möglichst genau beschrieben werden.
Die Herausforderung in der Bioinformatik ist, aus der dabei entstehenden riesigen Menge an Informationen über Gene, Proteine und Stoffwechsel- und Regulationsprozesse ein Verständnis der Funktionsweise einer Zelle bzw. eines kompletten Organismus zu entwickeln, d.h. zu verstehen, wie Gene und Biomoleküle zusammenwirken (Abbildung 1).

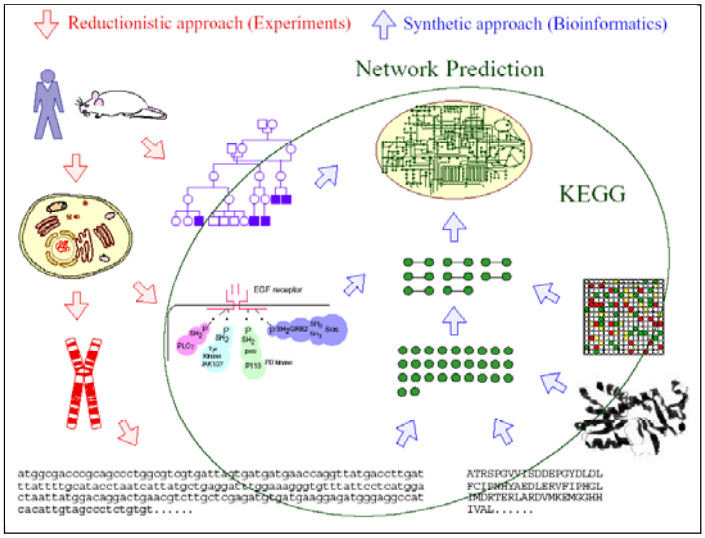


Abb. 1: Reduktionistische vs. Systembetonte Sichtweise [Kan01]

Für diese Aufgabe der sich gerade entwickelnden Systembiologie [BM02] reicht die genaue Kenntnis der Einzelkomponenten eines biologischen Systems nicht aus, vielmehr gilt es, biologische Prozesse auf der Systemebene zu untersuchen, indem die komplexen Netzwerke, welche die komplexe Funktionalität ausmachen, analysiert werden. Dazu ist es auch erforderlich, einen theoriebasierten Zugang zu der intuitiv nicht mehr fassbaren Komplexität zu finden. Die mathematische Modellierung ist hierbei ein wichtiges Werkzeug.

Damit die gewonnen Daten für systembiologische Forschung genutzt werden können, ist die Verwaltung und Bereitstellung der Daten in geeigneten Datenbanken unabdingbar. Neben den Sequenzdatenbanken, die Genom- bzw. Aminosäuresequenzen enthalten, sind für die Systembiologie jedoch vor allem neuere Datenbankgenerationen von Bedeutung, die beispielsweise Wechselwirkungen zwischen Proteinen und DNA, Stoffwechselwege oder Informationen zur Genexpression und -regulation enthalten.
Die vorliegende Arbeit beschäftigt sich mit den zu dieser Gruppe zählenden Datenbanken biochemischer Reaktionswege, sogenannten Pathway-Datenbanken. Zum besseren Verständnis der sehr komplexen Netzwerke biochemischer Reaktionen und ihrer Visualisierung und Repräsentation in Pathway-Datenbanken behandelt das folgende Kapitel wichtige Bausteine solcher Gen- und Chemieuniversen. In Kapitel 3 werden Pathway-Datenbanken im Allgemeinen und ihre praktische Anwendung anhand zweier Vertreter, KEGG und EcoCyc, beschrieben.

Abbildung 2 zeigt Ausschnitte aus dem 'Biochemical Pathways' Poster der Firma Boehringer Mannheim ([BMBPC]), dessen digitalisierte Version unter [ExPASy-Pathways] bereits eine einfache Pathway-Datenbank darstellt. Darin sind wesentliche biochemische Reaktionen in Zellen abgebildet. Die auf dem Poster zu den metabolischen Pfaden dargestellten etwa 1500 Reaktionen und ähnlich viele durch die Reaktionen umgesetzte Substanzen stellen einen großen Teil des Wissens über Stoffwechselwege zu Anfang der 90er Jahre dar. Heute sind schätzungsweise zehntausend Reaktionen und Substanzen bekannt [Sc01]. Doch auch dies ist nur ein kleiner Teil aller in Zellen ablaufenden Reaktionen. Allein das Genom des Menschen wird

von Molekularbiologen auf 30000 bis 45000 Gene geschätzt. Viele Gene kodieren ein oder mehrere Enzyme und jedes Enzym katalysiert eine oder mehrere Reaktionen. Konservative Schätzungen gehen davon aus, dass typische Zellen höherer Organismen einige Zehntausend verschiedene Proteine synthetisieren. Bahnsen [Ba00] verweist auf Schätzungen, nach denen durch die Information im menschlichen Genom sogar bis zu 20 Millionen verschiedene Proteine produziert werden. Die bisher aufgeklärten Reaktionen stellen also nur einen winzigen Bruchteil aller biochemischen Reaktionen dar. Dieses Wissen wächst jedoch täglich.

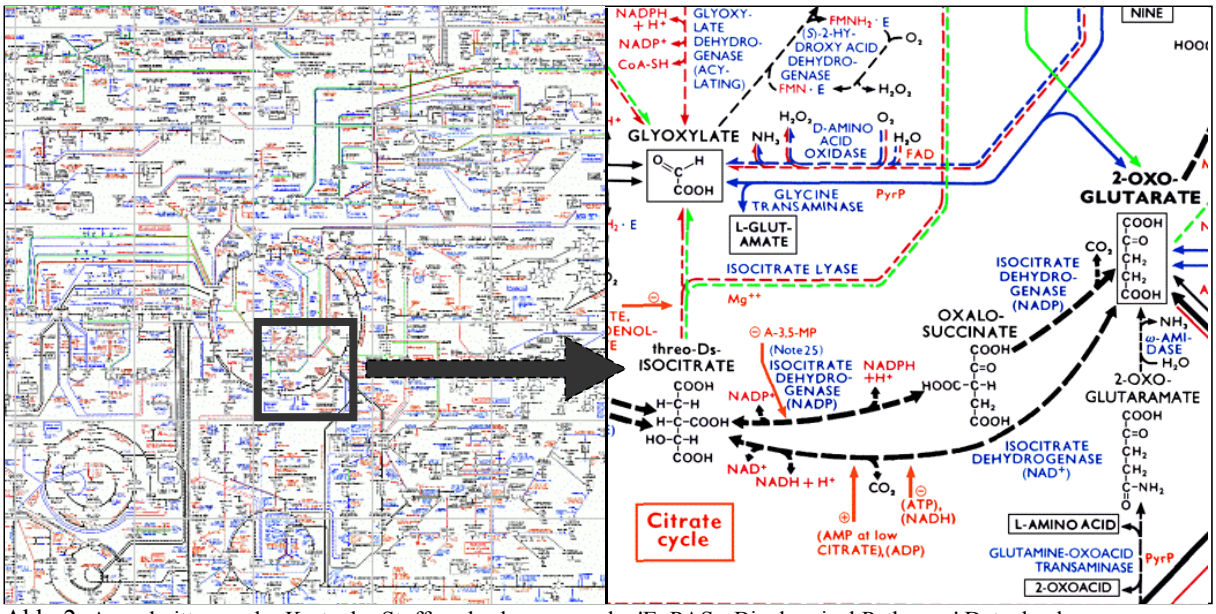


Abb. 2: Ausschnitte aus der Karte der Stoffwechselwege aus der 'ExPASy-Biochemical Pathways' Datenbank. [ExPASy-Pathways] Namen bezeichnen Substanzen und Enzyme, die Pfeile zwischen den Substanzen symbolisieren überwiegend deren Umsetzung durch Reaktionen. Ein elektronischer Index erleichtert die Suche nach einem Zwischenprodukt oder einem Enzym auf der Karte, den meisten Enzymen ist zusätzlich ein Hyperlink zur [ExPASy-ENZYME] Datenbank zugeordnet.

Pathway-Datenbanken veranschaulichen die sehr komplexen Netzwerke der Lebensvorgänge, indem sie graphische Abbildungen (Pathway-Maps) zu den biochemischen Reaktionswegen erzeugen oder enthalten.
Eine wichtige Möglichkeit der Anwendung von Pathway-Datenbanken ist es jedoch, biologische Theorien (wie z.B. das Stoffwechselnetz des E.coli-Bakteriums) derart strukturiert zu beschreiben, dass mit Hilfe von entsprechenden Analyseprogrammen das Rechnen mit diesen komplexen Daten zu neuen Erkenntnissen und Schlussfolgerungen führen kann [Kar01]. Die Vorhersage von Reaktionspfaden aus den bekannten Gensequenzen eines Organismus ist so eine Anwendung der Pathway-Analyse, deren Verfahren in dieser Arbeit aber nur am Rande erwähnt werden.

Wenn man weiß, wie bestimmte Lebensprozesse auf molekularer Ebene und darüber hinaus ablaufen und geregelt werden, eröffnet dies u.a. neue Möglichkeiten und Ansätze für die Entwicklung von Medikamenten und Therapien gegen Krankheiten. So ist beispielsweise mit Hilfe einer Pathway/Genom-Datenbank und zugehörigen Analysewerkzeugen eine Eingrenzung und Bestimmung effektiver Ansatzpunkte für antibakterielle Arzneistoffe möglich [KK+99].

2. Pathways und biologische Grundlagen

Leben kann auf unterster (molekularer) Ebene als eine Menge von biochemischen Prozessen und chemischen Reaktionen verstanden werden, wozu jeder Vorgang zählt, bei dem Moleküle miteinander interagieren und eine chemische oder physikalische Veränderung des Systems bewirken.
Biochemische Reaktionen finden in Zellen statt, wo sie die Bausteine der Zellen, also chemische Elemente und Verbindungen, umbauen. Eine biochemische Reaktion beschreibt die Umwandlung eines oder mehrerer Stoffe (Edukte oder Reaktanten) unter Veränderung der chemischen Bindungen zwischen Atomen zu einem oder mehrerer anderer Stoffe, den Reaktionsprodukten. Die Erzeugung von Energie, die Synthese von Substanzen, Wachstum, Vermehrung und Reaktion auf Umwelteinflüsse sind an biochemische Reaktionen gebunden. Systeme interagierender Proteine bilden hierbei die Grundlage für nahezu jeden Prozess in Lebewesen.

Als Pathway bezeichnet man in diesem biologischen Kontext eine Menge von zusammenhängenden biochemischen Reaktionen, wie sie in einer Zelle oder einem Organismus stattfinden. Zusammenhängend bedeutet: Das Reaktionsprodukt der einen Reaktion ist Reaktionspartner einer Folgereaktion oder aber ein Enzym, das eine weitere Reaktion katalysiert. Der spezielle Prozess, der durch den Pathway abgebildet wird, ist meist für bestimmte Aufgaben im Organismus zuständig. Erstaunlich viele dieser biochemischen Reaktionswege sind in allen Organismen vorhanden (z.B. bestimmte grundlegende Stoffwechselwege wie die Glycolyse) und ähnlich in ihrem Ablauf.
Biochemische Reaktionsnetze als nächsthöhere Stufe der Betrachtung sind beliebige Ausschnitte aus der Gesamtheit aller Reaktionswege. Sie lassen sich als Graphen modellieren, wobei die Knoten des Graphen die Substanzen oder Proteine symbolisieren und die Kanten die zwischen den Molekülen stattfindenden Interaktionen.

Hinsichtlich des Typs von Interaktionen lassen sich Pathways in zwei grundsätzliche Gruppen unterteilen:

- Stoffwechselwege oder metabolische Pfade (metabolic pathways):
  Sie spiegeln die Abläufe des Zellstoffwechsels wieder, die insbesondere durch Enzym-gesteuerte Reaktionen gekennzeichnet sind.
- Regulatorische Pfade (regulatory pathways):
  Sie beschreiben vor allem die Mechanismen der Genexpression und ihrer Regulation sowie Signalwege, Transportsysteme und andere zelluläre Prozesse.

Im folgenden werden nur einige wichtige Aspekte zu den grundsätzlich verschiedenen Pfadtypen genannt. Am Beispiel der Glycolyse als grundlegender Stoffwechselweg und dem p53-Signalweg als wichtiger Mechanismus im Zellzyklus werden diese verdeutlicht. Für eine umfassendere Darstellung der biochemischen Grundlagen sei auf [RH01], [Hu93], [Mi99] verwiesen.

2.1 Metabolische Pfade

Stoffwechselvorgänge sind historisch gesehen schon sehr früh untersucht worden. Das ist auch ein Grund dafür, weshalb die Forschung auf diesem Gebiet schon relativ viele und detaillierte Erkenntnisse gebracht hat.
Die genaue Kenntnis der Stoffwechselprozesse ist beispielsweise nützlich zur Diagnose und Behandlung von Stoffwechsel-Krankheiten, die u.a. durch fehlende Enzyme verursacht sein

können (Beispiel Fructose-1,6-diphosphatase-Mangel). Aber auch die Erforschung neuer Angriffspunkte für Medikamente, um gezielt den Stoffwechsel von Krankheitserregern oder Tumorzellen zu hemmen, benötigt detailliertes Wissen über die metabolischen Abläufe.

2.1.1 Stoffwechsel und Enzyme

Die Gesamtheit aller lebensnotwendigen biochemischen Vorgänge beim Aufbau, Abbau und Umbau eines Organismus bzw. beim Austausch zwischen dem Organismus und seiner Umwelt wird als Stoffwechsel (Metabolismus) bezeichnet. Die zugehörigen chemischen Reaktionen laufen sowohl in Pflanzen, Tieren und Menschen, als auch in Mikroorganismen ab und dienen entweder dem Aufbau und der Speicherung von Körper- bzw. Zellsubstanz (Anabolismus, Assimilation), oder ihrem Abbau (Katabolismus, Dissimilation). Diese entgegengesetzten Stoffwechselvorgänge sind eng miteinander verknüpft und werden hauptsächlich über den aktuellen Vorrat und Bedarf der Zellen an ATP (Adenosintriphosphat) als deren Energiewährung geregelt.

Fast alle biochemischen Reaktionen des Stoffwechsels werden durch Enzyme katalysiert. Diese sogenannten Biokatalysatoren sind Biomoleküle, meist Proteine, die die notwendige Aktivierungsenergie herabsetzen und die Reaktion beschleunigen. Dazu bildet das Enzym mit dem umzusetzenden Stoff, dem Substrat, einen Komplex, sodass die Reaktion unter den Zellbedingungen ablaufen kann. Das Substrat wird hierbei zum Reaktionsprodukt umgewandelt, das Enzym selbst bleibt jedoch unverändert.
Enzyme wirken substratspezifisch: ein bestimmtes Enzym setzt nur bestimmte Substrate zu bestimmten Produkten um. Hochspezifische Enzyme setzen sogar nur ein Substrat um. Beispielsweise katalysiert das Enzym Hexokinase die Umsetzung von Hexosen, wie Glucose oder Fructose, mit ATP zu Hexose-6-phosphat (Phosphorylierung), das Enzym Glucokinase dagegen phosphoryliert nur Glucose (niedrige vs. hohe Substratspezifität).

Enzyme bestimmen also aufgrund ihrer Spezifität, welche Stoffwechselreaktionen in der Zelle effizient ablaufen können. Die katalytische Aktivität von Enzymen hängt von verschiedenen Faktoren ab: neben der Enzymmenge, der Substratkonzentration und oftmals der Konzentration des Produktes (Produkthemmung) sind auch andere Moleküle in der Lage, die Enzymaktivität zu beeinflussen. Viele dieser Substanzen, körpereigene und körperfremde, hemmen die Aktivität. Man unterscheidet kompetitive Inhibitoren, die mit dem Substrat um die Bindungsstelle des Enzyms konkurrieren, und nichtkompetitive Inhibitoren. Auch ist irreversible Hemmung möglich, bei der der Inhibitor dauerhaft an die Substratbindungsstelle bindet und somit das aktive Zentrum des Enzyms irreversibel blockiert. Beispielsweise blockiert Penicillin die Synthese der Bakterienzellwand, indem es irreversibel mit der bakteriellen Glycopeptid-Transpeptidase reagiert. Die Transpeptidase kann dann die Polypeptidketten der Zellwand nicht mehr vernetzen. Penicillin reagiert nur mit dem bakteriellen und keinem menschlichen Enzym. Diese Spezifität lässt seinen Einsatz als Antibiotikum beim Menschen zu.
Sogenannte allosterische Enzyme werden durch Moleküle reguliert, die nicht im aktiven Zentrum sondern an einer räumlich distinkten Stelle binden (allosterische Effektoren). Allosterische Enzyme treten häufig an Verzweigungsstellen des Stoffwechsels auf. Durch Feedback-Hemmung, wobei das allosterische Enzym am Anfang eines Stoffwechselweges sitzt und von dessen Endprodukt gehemmt wird, ist eine effiziente Regulation von Stoffwechselwegen möglich (Vermeidung der Anhäufung von Zwischenprodukten).
Einige Enzyme benötigen für ihre Aktivität Cofaktoren. Dabei wird unterschieden zwischen solchen, die kovalent an das Enzym gebunden sind (prosthetische Gruppen) und denen, die nur vorübergehend mit dem Enzym in Wechselwirkung treten (Coenzyme). Wichtige Cofaktoren sind z.B. ATP, NAD^+, Coenzym A.

In vielen Fällen können enzymatische Reaktionen in beiden Richtungen ablaufen (reversibel), wobei sie durch ein und dasselbe Enzym katalysiert werden. Die Richtung der Umsetzung hängt dabei u.a. von der Substratkonzentration ab.

Für die Beschreibung und Einordnung der Enzyme und der von ihnen katalysierten Reaktionen existiert ein standardisierter EC-Code (Enzyme Commission, [IUBMB]). Die hierarchische Nummerierung der EC-Nomenklatur teilt die Enzyme in verschiedene Klassen ein. Die erste Kennnummer beschreibt die chemische Klasse des Enzyms bzw. der Reaktion: Oxidoreduktase(1), Transferase(2), Hydrolase(3), Lyase(4), Isomerase(5), oder Ligase(6). Kennnummer zwei und drei unterscheiden die Enzyme bezüglich ihrer Substrateigenschaften und anderer Reaktionsbedingungen in ihrer jeweiligen Unterklasse. Die vierte Ziffer ist die laufende Nummer innerhalb der untersten Ebene, die letztendlich für eine Gruppe von Proteinen steht, die die gleichen katalytischen Eigenschaften besitzen.

Bei metabolischen Pfaden stehen also Interaktionen von Enzymen untereinander und mit ihren Substraten im Vordergrund der Betrachtung. Den Kanten (Interaktionen) des Graphen eines metabolischen Reaktionsnetzes entsprechen in diesem Fall enzymatische Reaktionen. Die Verbindung zwischen zwei Substraten wird durch das Enzym hergestellt, das die entsprechende Reaktion katalysiert. Eine weitere Betrachtungsweise ist: Zwei Enzyme interagieren miteinander, indem das von dem einen Enzym katalysierte Produkt in der Folgereaktion das umzusetzende Substrat des zweiten Enzyms darstellt.

2.1.2 Glycolyse als grundlegender Stoffwechselweg

Glycolyse heißt die Serie von enzymatischen Reaktionen im Cytoplasma, die den C6-Zucker Glucose zu zwei C3-Einheiten (Pyruvat) abbaut.
Dieser Stoffwechselweg wurde mit als erster intensiv untersucht und ist einer der am besten verstandenen, was die beteiligten Enzyme, deren Wirkungsmechanismus sowie die Regulation des Stoffwechselweges mit einschließt. Er ist evolutionär gesehen sehr alt und in allen Organismen vorhanden [BioCarta].

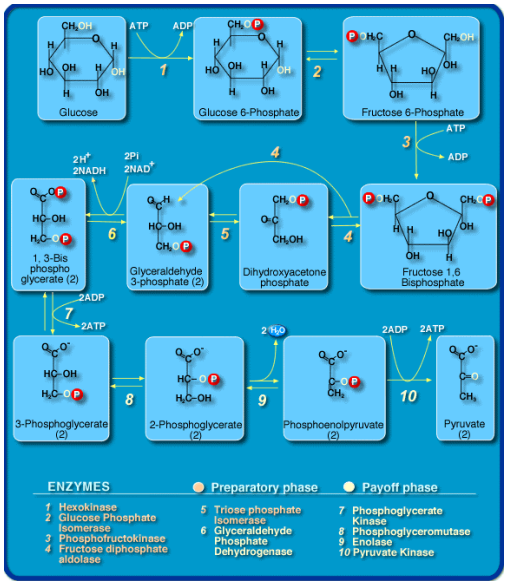


Abb. 3: Glycolyse beim Menschen, [BioCarta]

Die Glycolyse hat zwei Aufgaben: ATP zu produzieren und Vorstufen für Biosynthesen zu liefern (z.B. Pyruvat, woraus Acetyl-CoA für die Fettsäuresynthese entsteht). Regulatorisches Enzym der Glycolyse ist die Phosphofructokinase, sie bestimmt die Geschwindigkeit des Pathways und ist sein Kontrollpunkt. Das Enzym katalysiert Schritt 3 der Glycolyse (siehe Abbildung 3), die Phosphorylierung von Fructose-6-phosphat. Wie auch am E.C.-Code 2.7.1.11 erkennbar, gehört es zur Klasse der Transferasen, genauer zu den Phosphotransferasen (2.7) mit alkoholischer Gruppe als Akzeptor (2.7.1). Den Aufgaben der Glycolyse gemäß hemmen hohe ATP-Konzentrationen das Enzym, AMP aktiviert es. Des weiteren wird die Phosphofructokinase von Citrat gehemmt, denn hohe Citrat-Konzentrationen zeigen an, dass genügend Vorstufen für Biosynthesen vorhanden sind.
Endprodukt der Glycolyse ist Pyruvat, das anschließend drei verschiedene Reaktionswege einschlagen kann: Citratzyklus, Gluconeogenese, oder Reduktion zu Lactat.

2.2 Regulatorische Pfade

Neben den biochemischen Reaktionen, die den Stoffwechsel ausmachen, spielen regulatorische Mechanismen wie die Regulation der Genexpression und die Signaltransduktion eine entscheidende Rolle bei den Lebensprozessen. Hier werden Protein-Gen sowie Protein-Protein Wechselwirkungen betrachtet, die durch eine ganze Reihe von verschiedenen Vorgängen oder chemischen Reaktionen bestimmt sein können, beispielsweise durch Aktivierung oder Hemmung der Transkription eines Gens, Aktivierung oder Inaktivierung von Proteinen, Signalübertragung (Phosphorylierung) oder Transportvorgänge.

2.2.1 Regulation der Genexpression

Von der Genregulation hängt es ab, welche Proteine gerade in der Zelle synthetisiert und somit für Interaktionen wie Stoffwechsel verfügbar werden. Bestimmte DNA-bindende Proteine, sogenannte Transkriptionsfaktoren steuern die Initiation von Transkriptionsvorgängen. Einige Gene codieren Proteine, deren Aufgabe es ist, andere Gene in ihrer Expression an- oder abzuschalten. Viele dieser Regulationsproteine binden dazu an bestimmten Kontrollregionen in der Umgebung der codierenden DNA-Abschnitte, wo sie dann verstärkend oder hemmend auf die Genexpression wirken.
Bestimmte Gruppen von Regulationsgenen bilden komplexe Netzwerke zur Steuerung der Expression von Genen, deren Genprodukte bestimmte biochemische Reaktionen katalysieren. Die daran beteiligten Moleküle als Substrat oder Produkt der Reaktion können dann wiederum Regulationsproteine aktivieren oder deaktivieren. Die biochemischen Prozesse des Stoffwechsels und die Regulation zellulärer Prozesse sind also stark miteinander vernetzt.

2.2.2 Der p53-Signalweg

Eine menschliche Zelle wird zur Krebszelle, wenn die Kontrolle über ihr Teilungsverhalten verloren geht. Die Zelle reagiert nicht mehr auf wachstumshemmende Signale und durchbricht die Grenzen zwischen den Organen: sie metastasiert. Der Umwandlung einer normalen Körperzelle in eine Krebszelle liegen Mutationen in Protoonkogenen und Tumorsuppressor-Genen zugrunde. p53 ist solch ein Tumorsuppressor-Gen.
Das Eiweiß P53 wurde 1979 unabhängig voneinander von dem schottischen Forscher David Lane und von dem Amerikaner Arnold Levine entdeckt. Wenige Jahre später fand man auch das entsprechende Gen, dachte aber, man habe es wieder mit einem der vielen bereits bekannten Onkogene zu tun. Erst als man 10 Jahre später das wahre Gesicht von p53 erkannte, stieg das Interesse der Wissenschaftler: Man fand heraus, dass es nicht, wie man vorher glaubte, eine gesunde Zelle zur Tumorzelle umwandelt, sondern im Gegenteil normale Zellen vor der Entartung bewahrt. Seither ist p53 ins Zentrum der Forschung gerückt. Es wird auch als Wächter des Genoms bezeichnet.
Das p53-Netzwerk in Abbildung 4 verdeutlicht die zentrale Rolle des Transkriptionsfaktors p53 bei der Kontrolle des Zellzyklus. p53 ist der zentrale Knoten des Netzwerkes, der die meisten Ein- und Ausgänge besitzt. Es ist daher nicht überraschend, dass ein Ausfall der ursprünglichen Funktion von p53 diesen für die geregelte Zellteilung so wichtigen Signalweg erheblich stört. So konnte in mehr als 50% aller menschlichen Tumore eine Mutation von p53 festgestellt werden.

Das p53-Netzwerk, wie in Abbildung 4 dargestellt, ist normalerweise ausgeschaltet. Erst durch schädigende Einflüsse auf die Zelle wird es aktiviert. Einer dieser Auslösefaktoren ist die Schädigung von DNA, wie sie z.B. durch Röntgenstrahlung hervorgerufen wird. DNA-Brüche bewirken eine Aktivierung von mehreren Protein-Kinasen (DNA-abhängige Kinase, ATM u.a.), Enzyme, die Phosphatgruppen auf andere Proteine übertragen. Das phosphorylierte p53 ist dann nicht mehr in der Lage, die Transkription des Proteins MDM2 zu stimulieren, das für den Abbau von p53 verantwortlich ist. (feedback loop). Dadurch kommt es zu einem Anstieg der Konzentration an aktiviertem p53 im Zellkern.

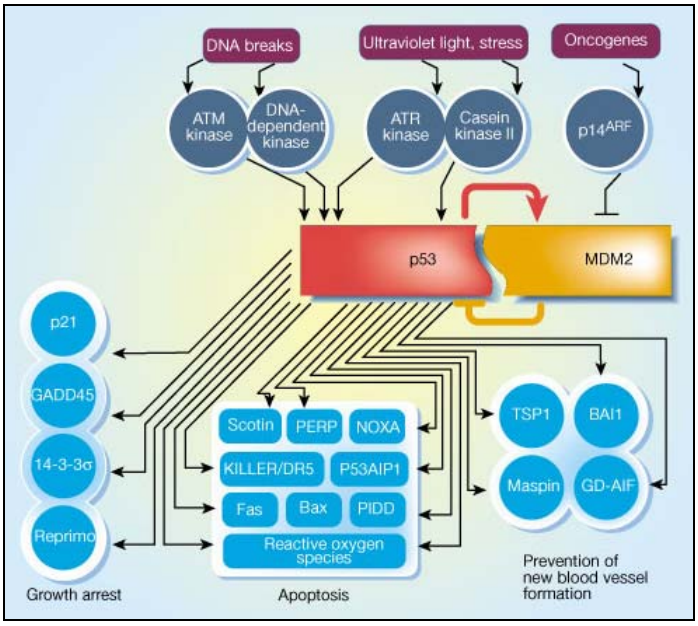


Abb. 4: Das p53-Netzwerk, [VL+00]

Aktiviertes p53 kann dann als Transkriptionsfaktor aktiv die Expression von bestimmten Genen stimulieren, die unter anderem ein Fortschreiten des Zellzyklus unterbinden (Inhibitorprotein p21, Reprimo u.a.), bis eine Reparatur erfolgt ist, oder auch den programmierten Zelltod (Apoptosis) einleiten (Bax-Protein, NOXA, Fas u.a.).

Wie auch in [VL+00] betont wird, können Signalwege und -netze im Allgemeinen wie auch bei p53 nicht verstanden werden, indem nur die Einzelkomponenten betrachtet werden. Es ist notwendig, das komplizierte Netzwerk aus diesen signalgebenden Komponenten zu untersuchen. Eine wichtige Frage in der Krebs-Forschung ist es beispielsweise, wie ein zelluläres Netzwerk gezielt attackiert werden kann, das bereits an zentraler Stelle (p53) in der ursprünglichen Funktion gestört ist (Tumorzelle).

3. Pathway-Datenbanken und ihre Anwendung

Mehrere webbasierte Dienste bieten Zugang zu Informationen über Stoffwechsel- und andere Reaktionswege. Diese Ressourcen sind hauptsächlich Datenbanken mit Suchmechanismen und Verknüpfungen zu weiteren Quellen.
Biologische Datenbanken, die biochemische Reaktionswege und deren Einzelreaktionen, beteiligte Enzyme und Substrate beschreiben, werden als Pathway-Datenbanken bezeichnet.

Enzym-Datenbanken, wie [ExPASy-ENZYME] oder [BRENDA], enthalten hauptsächlich Informationen über Enzyme und ihre speziellen Eigenschaften, und beschreiben damit auch biochemische Reaktionen. Die komplexeren Pathway-Datenbanken geben Informationen zu Reaktionen und Reaktionswegen im Allgemeinen, organismenspezifische Daten über Gene und zugehörige Genprodukte, Proteinfunktionen, Expressionsdaten etc.
Pathway-Datenbanken stehen aber in den meisten Fällen eng in Verbindung mit Enzym-Datenbanken. Abbildung 5 gibt einen Überblick über die Verknüpfungen wesentlicher Datenbanktypen der Biologie. Ein Hauptvorteil von Pathway-Datenbanken gegenüber anderen biologischen Datenbanken ist, dass sie die verschiedenen Informationen im Gesamtzusammenhang in Form graphischer Pathway-Darstellungen bereitstellen.

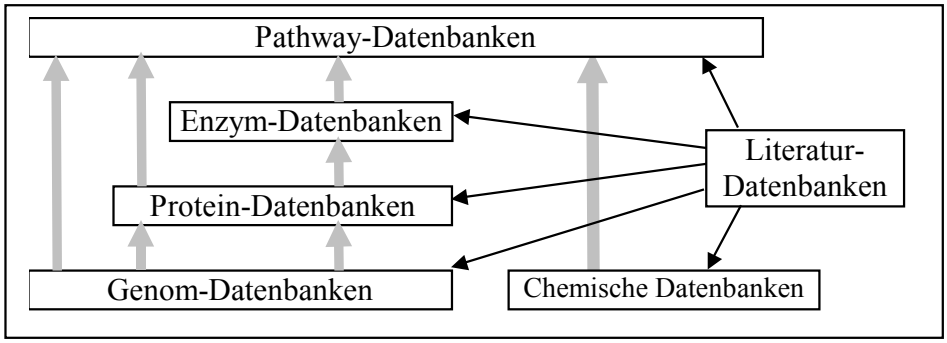


Abb. 5: Die Verknüpfungen unter den wesentlichen Datenbanktypen, nach [WD01]

Wichtige Eigenschaften zur Einordnung von Pathway-Datenbanken sind [Sh02]:

- *die interne Repräsentation der Pathway-Daten*:
  Viele sogenannte Datenbanken beinhalten anschauliche, graphische Pathway-Abbildungen, aber außer einfachen Such- und Abfragemöglichkeiten und der Verknüpfung von Einträgen sind oftmals keine komplexeren Anfragen oder das Rechnen mit den Pathway-Informationen möglich.
  Beispiele: [BioCarta], [ExPASy-Pathways]

  Hingegen erlauben Pathway-Datenbanken mit strukturierter interner Repräsentation der Daten, z.B. in Form von Graphen, komplexere Anfragen und bieten meist entsprechende Tools bzw. Schnittstellen zur Analyse und Berechnung auf den Pathway-Informationen an.
  Beispiele: [KEGG], [MetaCyc], [EcoCyc], [WIT]

- *der Pathway-Typ: metabolische und/oder regulatorische Pfade:*
  Die Mehrheit der zur Zeit existierenden Datenbanken konzentriert sich auf metabolische Pathways, so z.B. [MetaCyc], [KEGG], [MALARIA], [WIT].
  Pathway-Datenbanken speziell zu regulatorischen Pfaden und Regulationsnetzwerken sind beispielsweise [CSNDB], [SPAD].

- *die Anzahl beschriebener Organismen:*
  Einige Pathway-Datenbanken haben sich auf spezielle Organismen oder Organismengruppen spezialisiert, wie z.B. [EcoCyc] auf Escherichia coli, oder [MALARIA] auf Malaria-Parasiten. Viele enthalten aber Informationen bezüglich mehrerer Arten ([KEGG], [MetaCyc]).

- *Einbeziehung statistischer Unsicherheit:* Die meisten Datenbanken berücksichtigen den Aspekt der Unsicherheit der Daten bzw. Verlässlichkeit der übernommenen und erzeugten Daten nicht ausreichend.

In den folgenden Abschnitten werden, nach einem Überblick zu den derzeit aktuellen Datenbankprojekten (Abschnitt 3.1), zwei wichtige Pathway-Datenbanken, KEGG und EcoCyc, in ihrem Aufbau und den Anwendungsmöglichkeiten näher beschrieben.

3.1 Überblick zu aktuellen Pathway-Datenbanken

Es existieren eine ganze Reihe von Datenbankprojekten zu Pathway-relevanten Informationen, die sich, wie schon am Beginn des Kapitels gezeigt, in ihren grundsätzlichen Merkmalen stark unterscheiden können.
Die aktuellen Projekte der Kategorie 'Metabolic Pathways and Cellular Regulation' aus dem Nucleic Acids Research Sonderheft vom Januar 2003 [Ba03] sind in Tabelle 1 aufgeführt. Sie geben einen Überblick über die zur Zeit in der Forschung relevanten und bekanntesten Pathway-Datenbanken.

EcoCyc	http://ecocyc.org/	*Escherichia coli* K-12 genome, metabolic pathways, transporters and gene regulation
ENZYME	http://www.expasy.ch/enzyme/	Enzyme nomenclature
EpoDB	http://www.cbil.upenn.edu/EpoDB/	Genes expressed during human erythropoiesis
Klotho	http://www.ibc.wustl.edu/klotho/	Collection and categorization of biological compounds
KEGG	http://www.genome.ad.jp/kegg	Metabolic and regulatory pathways
LIGAND	http://www.genome.ad.jp/ligand/	Chemical compounds and reactions in biological pathways
MetaCyc	http://ecocyc.org/	Metabolic pathways and enzymes from various organisms
The University of Minnesota Biocatalysis Biodegradation Database	http://umbbd.ahc.umn.edu/	Curated information on microbial catabolismand related biotransformations
PathDB	http://www.ncgr.org/pathdb	Biochemical pathways, compounds and metabolism
PRODORIC	http://prodoric.tu-bs.de	Prokaryotic database of gene regulation and regulatory networks
RegulonDB	http://www.cifn.unam.mx/Computational_Genomics/regulondb/	*Escherichia coli* transcriptional regulation and operon organization
UM-BBD	http://umbbd.ahc.umn.edu/	Microbial biocatalytic reactions and biodegradation pathways
WIT2	http://wit.mcs.anl.gov/WIT2/	Integrated system for metabolic models

Tabelle 1: Aktuelle Datenbanken zu Pathway-Informationen, [Ba03]

3.2 KEGG: Kyoto Encyclopedia of Genes and Genomes

Die 'Kyoto Encyclopedia of Genes and Genomes' ([KEGG], [KG+02]) ist eine sehr umfangreiche, öffentlich zugängliche Datenbanksammlung zu Pathway relevanten Informationen, die zu den Angeboten des japanischen GenomeNet Datenbank Services [GenomeNet] zählt.
Das KEGG-Projekt, das 1995 unter Leitung von Minoru Kanehisa [Kanehisa Laboratory] gestartet wurde, integriert das gesamte Wissen in Bezug auf die molekularen Interaktionsnetzwerke biologischer Prozesse (PATHWAY Datenbank), das Wissen über den Bereich der Gene und Proteine (GENES/SSDB/KO Datenbanken) und die Welt der chemischen Verbindungen und Reaktionen (COMPOUND/REACTION Datenbanken).
Die Datenobjekte in den einzelnen Datenbanken werden intern als Graphen repräsentiert, und es sind verschiedene Berechnungsmethoden in Entwicklung, um bestimmte Eigenschaften der Graphen zu erkennen und anschließend mit entsprechenden biologischen Bedeutungen in Zusammenhang zu bringen. So können z.B. durch Ähnlichkeitsvergleiche bestimmter Graphen die für einen Pathway oder Komplex zuständigen Gene vorhergesagt werden.

Number of pathways	9,644	(PATHWAY database)
Number of reference pathways	218	(PATHWAY database)
Number of ortholog tables	83	(PATHWAY database)
Number of organisms	115	(GENOME database)
Number of genes	417,535	(GENES database)
Number of KO assignments	3,336	(KO database)
Number of KO candidates	20,474	(SSDB database)
Number of chemical compounds	10,458	(COMPOUND database)
Number of chemical reactions	5,278	(REACTION database)

Tabelle 2: Übersicht über die in KEGG gehaltenen Daten, [KEGG]

Tabelle 2 zeigt eine Übersicht zur Anzahl der zur Zeit (Release 25.0, Januar 2003) in KEGG gespeicherten Datenobjekte.
KEGG enthält alle bekannten Stoffwechselwege und eine Auswahl an regulatorischen Pathways und Transportmechanismen [WD01].

Die 3 wichtigsten, stark miteinander verknüpften Datenbanken des KEGG Systems sind:

LIGAND: beinhaltet die Informationen über chemische Verbindungen, Enzyme und Reaktionen.

PATHWAY: enthält die graphischen Darstellungen der Reaktionswege und Listen der Enzyme und Reaktionen, die an diesen beteiligt sind.

GENES: enthält Genkataloge aller vollständig sequenzierten Genome und einiger unvollständig sequenzierter Genome sowie Listen der Gene, die innerhalb eines bestimmten Pathways eine Rolle spielen.

Zusätzlich sind in KEGG viele Links zwischen den KEGG-Datenbanken untereinander und zu externen Datenbanken integriert, die durch das 'DBGET database retrieval system' und die 'LinkDB'-Datenbank (Abbildung 6) verwaltet werden.

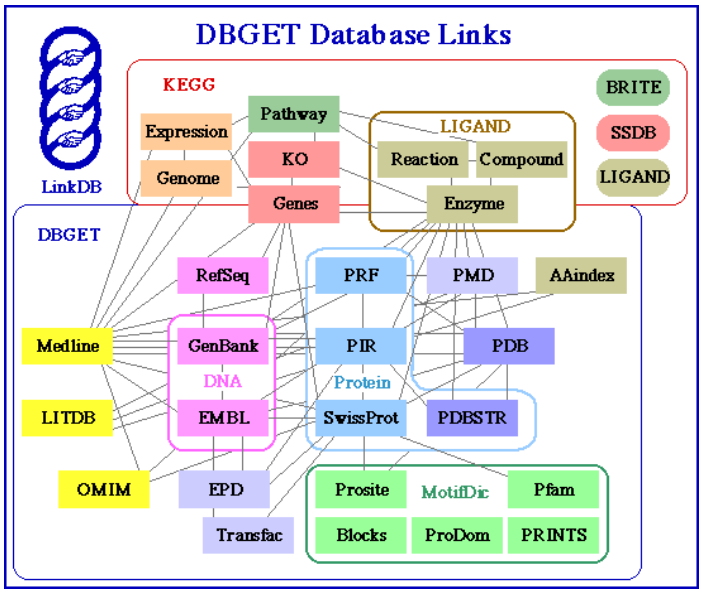


Abb. 6 : DBGET Integrated database retrieval system, [DBGET]

Die Daten der PATHWAY als auch LIGAND Datenbank wurden manuell aus der Literatur eingetragen, die Genkataloge in GENES stammen aus öffentlichen Sequenzdatenbanken wie GenBank und werden in ihren Annotationen ständig durch KEGG aktualisiert. Ziel dieser Annotationen ist es, die beschriebenen Gensequenzen in Verbindung mit orthologen Genbezeichnern (ortholog identifiers) zu bringen. Diese stehen für einander entsprechende Gene gleicher Funktion, die (auch aus den verschiedenen Organismen stammend) als Gene evolutionär verwandten Ursprungs zu betrachten sind.

Durch diese Bezeichner und ihre Erweiterung namens KEGG Orthology (KO) ist es möglich, aus den Genomdaten eines Organismus und einem allgemeinen Referenz-Pathway den organismenspezifischen Pathway zu bestimmen. Ein Referenzpathway enthält hierbei alle die Reaktionen und beteiligte Substanzen, die bei verschiedenen Organismen für einen bestimmten Reaktionsweg ermittelt werden konnten. Diese manuell erstellten Referenz-Diagramme bilden die Grundlage der KEGG/PATHWAY Datenbank. Die Integration mit den Genomdaten eines bestimmten Organismus führt dann zu organismenspezifischen Pathway-Maps. Das bedeutet z.B. im Fall von metabolischen Pfaden, dass nur diejenigen Enzyme eines Referenzpathways in den speziellen Pathway des Organismus übernommen werden, die auch in seinen Genen codiert werden. Dazu werden die entsprechenden Enzyme in den Pathway-Abbildungen farblich hervorgehoben. Abbildung 7 zeigt hierzu einen Ausschnitt aus dem Glycolyse-Pathway-Map des Menschen. Die automatische Verfahrensweise zur Vorhersage und Bestimmung von Pathways (Einsatz von Matching-Algorithmen auf Basis von Sequenzhomologien) ist möglich, da die verschiedenen Daten intern als Graphen repräsentiert werden.

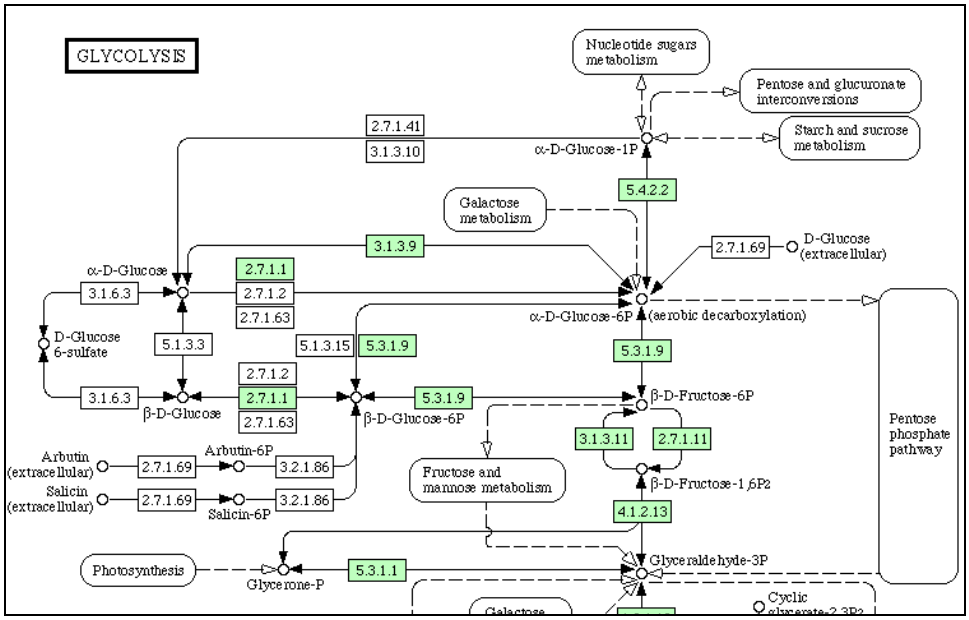


Abb. 7: Ausschnitt aus dem Pathway-Map zur Glycolyse/Gluconeogenese des Menschen, [KEGG]
Die experimentell oder durch Vorhersage bestimmten Enzyme sind grün unterlegt und mit organismenspezifischen Sequenzdaten verknüpft.

In KEGG existieren drei grundlegende Graphobjekte, die zur Darstellung und Manipulation von Daten verwendet werden. Diese und die zugehörigen Bausteine (nodes) und Interaktionen (edges) sind in Tabelle 3 abgebildet. Sie bilden die Grundlage der KEGG-Ontololgie.

Graph	Node	Edge	Database
Gene universe	Gene		GENES
	Protein-coding gene	Sequence similarity (orthology, paralogy, etc.)	SSDB
	Gene	Adjacency	GENES, GENOME
	Gene	Expression similarity	EXPRESSION
	Gene or gene product	Interaction or relation	BRITE
Protein network	Gene product or subnetwork	Direct protein-protein interaction Gene expression relation Enzyme-enzyme relation	PATHWAY
Chemical universe	Chemical compound		COMPOUND
	Chemical compound	Chemical reaction	REACTION

Tabelle 3: Modellierung der Daten in Form von Graphen, [KEGG]

Das 'Protein network' ist das wohl einzigartigste Graphobjekt in KEGG, das in Form von Pathway-Diagrammen in der PATHWAY Datenbank gespeichert ist. Wie in Tabelle 2 ersichtlich, gibt es darin 3 verschiedene Interaktionstypen (edges):

- Enzym-Enzym Relationen, die im Rahmen von Stoffwechselwegen aufeinanderfolgende katalytische Reaktionen symbolisieren
- direkte Interaktion zwischen zwei Proteinen, beispielsweise Phosphorylierung im Rahmen von Signalwegen
- Interaktionen zur Genexpression, an denen Transkriptionsfaktoren und Genprodukte teilnehmen

Das verallgemeinerte 'protein interaction network' wird manuell als graphisches Pathway-Diagramm erstellt (Referenzpathway) und zusätzlich als Menge binärer Relationen gespeichert. Die Modellierung als Menge binärer Relationen macht es wiederum möglich, mit diesem Informationsnetzwerk zu rechnen.

Die Einträge der PATHWAY Datenbank sind gemäß ihrer biologischen Funktion und der beteiligten Komponenten hierarchisch klassifiziert (Tabelle 4). In der ersten Ebene wird unterschieden zwischen Pathways bezüglich Stoffwechsel, Genexpression und -regulation, Interaktionen mit der Umgebung, zellulären Prozessen und menschlichen Krankheiten.
Über die Inhaltsübersicht (www.genome.ad.jp/kegg/kegg2.html) gelangt man am besten zum gesuchten Reaktionsweg, indem man sich durch diese Hierarchie navigiert.

Metabolism Carbohydrate Metabolism Energy Metabolism Lipid Metabolism Nucleotide Metabolism Amino Acid Metabolism Metabolism of Other Amino Acids Metabolism of Complex Carbohydrates Metabolism of Complex Lipids Metabolism of Cofactors and Vitamins Biosynthesis of Secondary Metabolites Biodegradation of Xenobiotics	Genetic Information Processing Transcription Translation Sorting and Degradation Replication and Repair Environmental Information Processing Membrane Transport Signal Transduction Ligand-Receptor Interaction Cellular Processes Cell Motility Cell Growth and Death Cell Communication Development Behavior Human Diseases Neurodegenerative Disorders

Tabelle 4: Hierarchische Struktur der Netzwerkinformationen in KEGG/PATHWAY, nach [KG+02]

Für den Zugang zu metabolischen Pfaden gibt es zusätzlich schematische Übersichtskarten, die die Navigation erleichtern (siehe Abbildungen 8 und 9). Alternativ besteht die Möglichkeit, über das schon erwähnte DBGET/LinkDB System nach Pathway-Einträgen zu suchen, oder ausgehend von Genkatalogen einzelner Organismen zu zugehörigen organismenspezifischen Pathways zu gelangen.

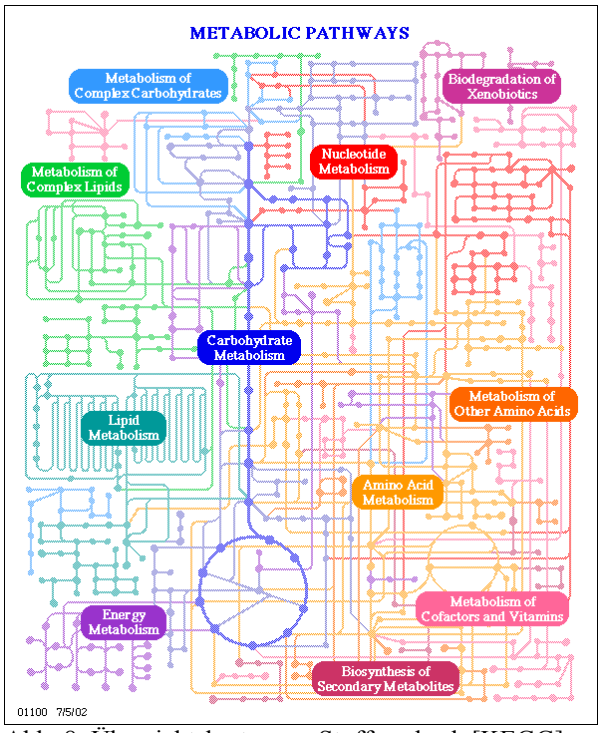


Abb. 8: Übersichtskarte zum Stoffwechsel, [KEGG]

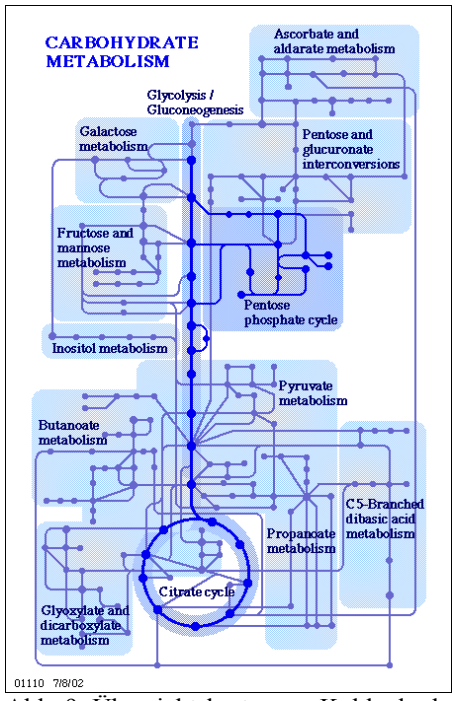


Abb. 9: Übersichtskarte zum Kohlenhydrat-Stoffwechsel, [KEGG]

Seit Januar 2003 sind die metabolischen Pfade auch im XML-Format verfügbar. Die Daten sind gemäß KGML (KEGG Markup Language) strukturiert, die durch eine DTD-Spezifikation beschrieben ist. Über ein Java-Applet (PathwayViewer) können die Stoffwechselwege aus den XML-Dateien zusätzlich dynamisch visualisiert werden. Gerade für den Datenaustausch und die Verwendung in eigenen Programmen bietet die Repräsentation in XML eine gute Möglichkeit der zusätzlichen Nutzung.

KEGG bietet neben den Pathway-Abbildungen zusätzliche Such- und Abfragemöglichkeiten, wie beispielsweise die Suche und farbliche Markierung von chemischen Verbindungen oder Enzymen in den Pathway-Maps, oder auch die Suche ähnlicher Sequenzen zu einer vorgegebener Sequenz innerhalb der Pathways.

Ein besonderes Feature des KEGG Systems ist es, mögliche Verbindungen (Pathways) zwischen zwei gegebenen chemischen Verbindungen bis zu einer bestimmten Länge (cut off length) berechnen zu lassen. Basierend auf den binären Relationen der Enzyme werden Algorithmen (Dijkstra/Floyd) zur Bestimmung des kürzesten Weges zwischen zwei Komponenten eingesetzt. Eine Einschränkung bezüglich des verwendeten Datensatzes (z.B. Homo sapiens) soll es ermöglichen, organismenspezifische Betrachtungen anzustellen. Das Ergebnis kann neben der textuellen Ausgabe (als Liste von EC-Nummern und IDs der chemischen Verbindungen) auch graphisch dargestellt werden, allerdings ist dieses bei den meist sehr komplexen Wegen und den verwendeten Abkürzungen (IDs) nur schwer zu verstehen.
Abbildungen 10 und 11 demonstriert dieses Feature an einem Beispiel.

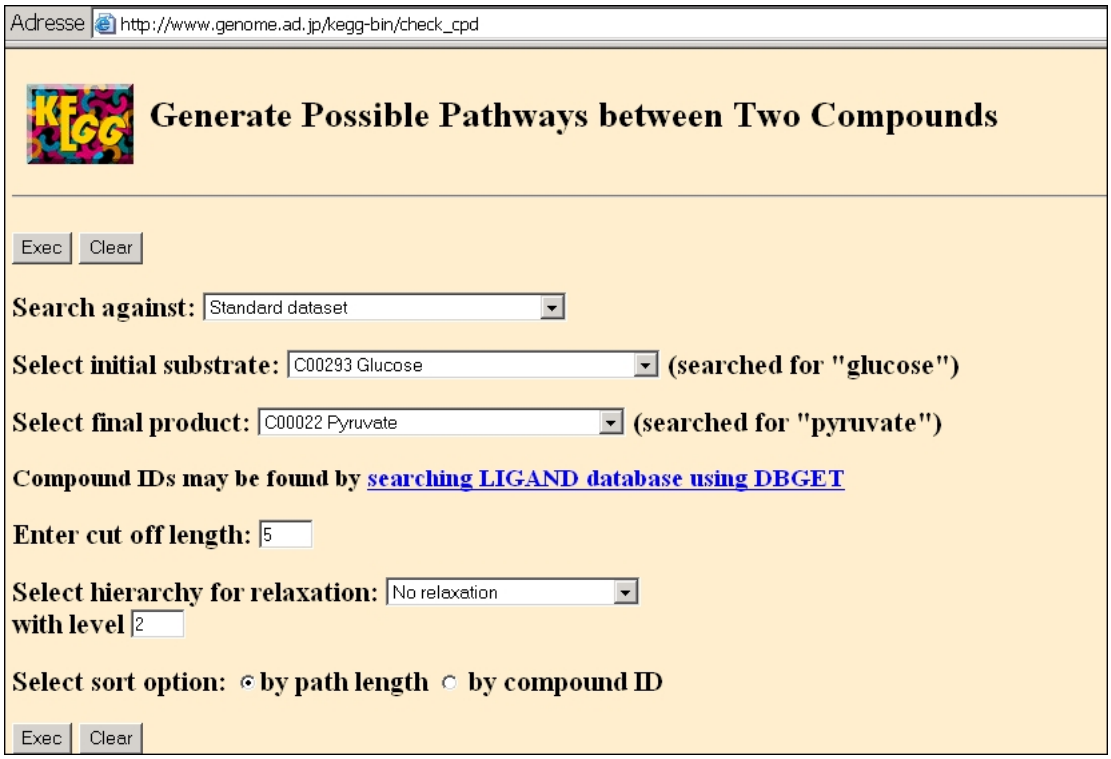


Abb. 10: Beispiel zur KEGG-Anwendung – Anfrageformular zur Berechnung aller theoretisch möglichen Pathways, die zwei Substrate (hier Glucose und Pyruvat) verbinden. [KEGG]

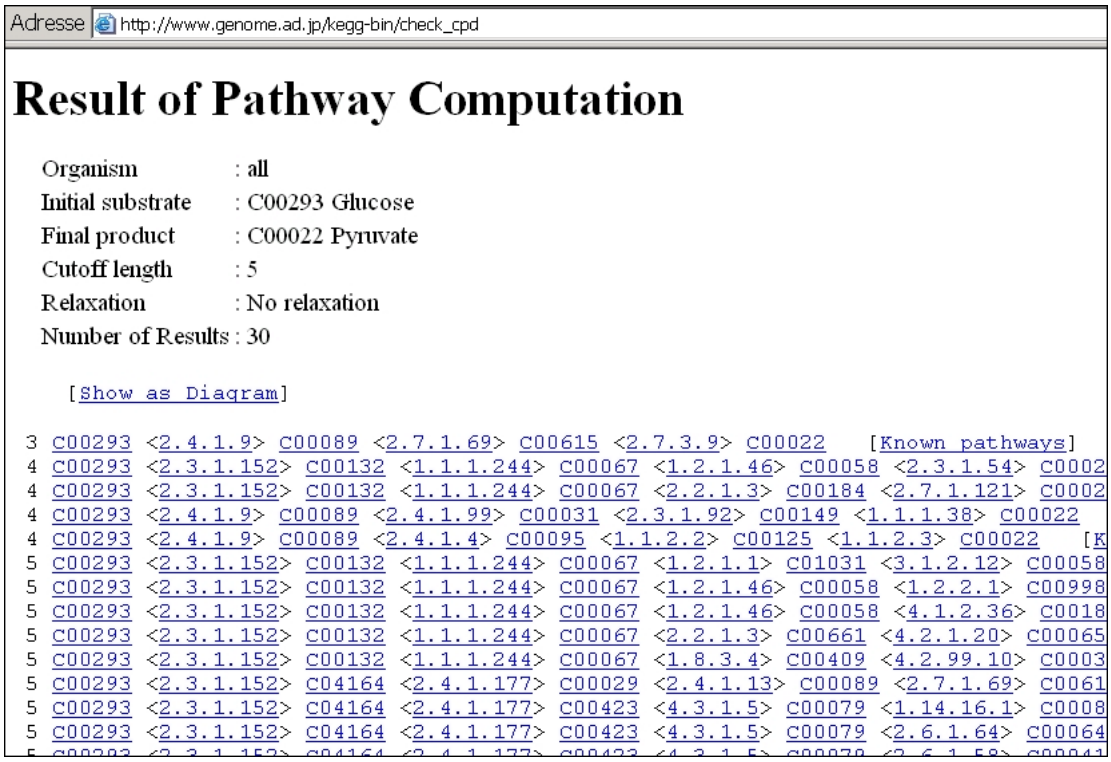
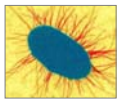


Abb. 11: Ergebnis der Berechnung anhand der Pathway-Daten (Standard Dataset) in [KEGG]: 30 Reaktionswege bei maximaler Länge von 5 Reaktionen. Eine Verdopplung der zulässigen Länge auf 10 führt bei dem verwendeten Standard-Dataset bereits zu über 37000 Ergebnissen.

Das gesamte KEGG-System kann außer über den Web-Zugriff auch lokal verwendet werden, indem die entsprechenden Daten samt Software installiert werden. (KEGG Distribution unter http://www.genome.ad.jp/kegg/kegg5.html)

3.3 EcoCyc: Encyclopedia of Escherichia coli Genes and Metabolism



EcoCyc ist eine organismenspezifische Pathway-Datenbank, genauer Pathway/Genom-Datenbank (PGDB). Sie enthält das gesamte Wissen über die biochemische Maschinerie des E.coli-Bakteriums zusammen mit dem bereits vollständig sequenzierten Genom.
Die Daten basieren auf den Einträgen der EcoGene Datenbank, SWISS-PROT und der wissenschaftlichen Literatur.
Zur 'BioCyc Knowledge Library' [BioCyc], einer öffentlich zugänglichen Sammlung von PGDBs unter Projektleitung von Peter D. Karp, gehören neben EcoCyc noch weitere Mikroorganismen-spezifische PGDBs, die in ihrer Struktur mit der von EcoCyc weitgehend übereinstimmen.
Die aktuelle EcoCyc-Statistik der Version 6.5 vom August 2002 ist in Tabelle 5 zu sehen. Bei der Zahl der Reaktionen sind neben Stoffwechselreaktionen auch Reaktionen im Zusammenhang mit Transportvorgängen und der Regulation der Genexpression auf Transkriptionsebene erfasst. Die EcoCyc-Datenbank beinhaltet Daten zu folgenden drei Gebieten:

EcoCyc KB Statistics	
Pathways	164
Reactions	2862
Enzymes	918
Transporters	168
Genes	4393
Transcription Units	724
Citations	3701

Tabelle 5: [EcoCyc]

E. coli Stoffwechsel: metabolische Pfade.
Jedes Stoffwechsel-Enzym sowie zugehörige Cofaktoren, Aktivatoren und Inhibitoren und Struktur der Untereinheiten sind erfasst.
E. coli Regulation der Transkription: E. coli Operons, Promoter, Transkriptionsfaktoren und zugehörige Bindungsstellen. EcoCyc enthält die vollständigste Beschreibung des genetischen Netzwerkes unter allen untersuchten Organismen
E. coli Membran-Transporter: Transportproteine und zugehörige Reaktionen.
E. coli Genom: vollständige Genomsequenz von E. coli MG1655, einschließlich Position und Funktion aller E.coli-Gene

Das EcoCyc System ist nach der sogenannten EcoCyc-Ontologie strukturiert. Die Ontologie (DB-Schema) beinhaltet ca. 1000 Klassen, die Schlüsselkonzepte der Biochemie und Molekularbiologie codieren, und mehr als 200 "Slots" (Attribute), die die Eigenschaften der Klassen und ihre Beziehungen untereinander charakterisieren. Die Ontologie erfasst somit wichtige semantische Unterschiede der relevanten Konzepte und definiert ihre Bedeutung innerhalb der Datenbank. Beispiele für Klassen sind 'Pathways', 'Compounds' oder 'Genes'. Die Klassen der Bereiche bezüglich Pathways, Reaktionen, Verbindungen und Gene sind jeweils gemäß einer Vererbungshierarchie strukturiert, d.h. eine Klasse besitzt mindestens eine Oberklasse (parent class) und/oder eine Menge an Unterklassen (child classes). So hat beispielsweise die Klasse 'Pathways' im Rahmen der Pathway-Klassifikation die Elternklasse 'Generalized-Reactions' und mehrere Unterklassen, wie z.B. 'Signal-transduction pathways' und 'Biosynthesis'. Die Slots einer Klasse beschreiben die Attribute, die jede Instanz (Objekt) einer Klasse besitzt. Jedes Pathway-Objekt besitzt z.B. den slot 'REACTION-LIST', der die beteiligten Reaktionen enthält.

Die EcoCyc Datenbank besteht aus einem Netz verknüpfter Objekte (Frames) gemäß der beschriebenen Ontologie, die in einem Wissensrepräsentationssystem (ähnlich einem objektorientierten Datenbanksystem) gespeichert sind [Kar01].

Der Datenzugriff wird über eine Software-Umgebung namens "Pathway Tools" [PTools] realisiert, die Anfrage-, Analyse- und Visualisierungsoperationen für EcoCyc und die anderen

PGDBs unterstützt. Visualisierungswerkzeuge generieren automatisch zur Laufzeit Abbildungen der Pathways, Reaktionen, chemischen Verbindungen, Chromosomen und Transkriptionseinheiten, die Ergebnisse einer Anfrage darstellen.
Die Navigation in visualisierten Pathway-Diagrammen wird erleichtert, indem zwischen verschiedenen Komplexitätsstufen der Darstellung gewechselt werden kann. Auf detailliertester Ebene werden neben den Reaktionen und den beteiligten Enzymen samt EC-Nummern und Cosubstraten u.a. auch die chemische Struktur der Substrate abgebildet. Abbildung 12 zeigt einen Ausschnitt aus dem Glycolyse-Pathway auf dieser Ebene.

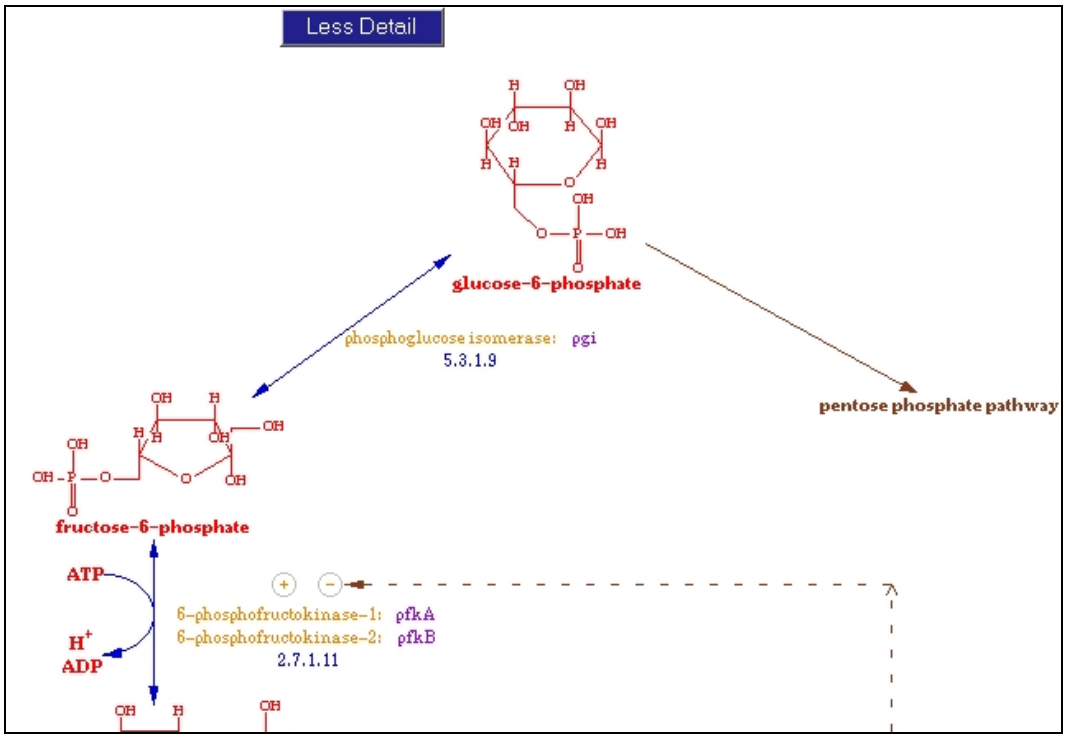


Abb. 12: Ausschnitt aus der Detailansicht des Glycolyse-Pathways von E. coli, [EcoCyc]

Verweise auf benachbarte Pfade (wie in Abbildung 12 der Pentose Phosphat Pathway) und Mechanismen der Regulation sind ebenfalls in die Darstellungen integriert. Jede Komponente des Pathway-Diagramms repräsentiert einen Link (DB-Anfrage) auf Detailinformationen des entsprechenden Objektes.

Suchanfragen an EcoCyc werden auf der 'BioCyc Query Page' über Formulare definiert. Neben einer einfachen Suche anhand von Bezeichnung oder EC Nummer der Komponenten in den verschiedenen Kategorien oder der Navigation durch die schon erwähnten Klassifikationshierarchien (Genes, Compounds, EC-Hierarchie, Pathways) besteht die Möglichkeit einer komplexeren Anfrage (Advanced Query Form). Die Suche innerhalb einer Kategorie kann hierbei durch Angabe von bis zu fünf Bedingungen an ausgewählte Slots präzisiert werden, die durch einen gemeinsamen logischen Funktor verknüpft werden. Als Ergebnis der Anfrage erhält man alle der Query-Bedingung genügenden Einträge zurück.

Einen anderen Zugang zu Pathway-Informationen in EcoCyc bietet das 'Metabolic overview diagram', eine schematische Übersichtskarte aller Stoffwechselwege der EcoCyc Datenbank (Abbildung 13). Graphische Symbole stellen die chemischen Verbindungen (Metabolite) des Reaktionsnetzes dar, die blauen Verbindungslinien stehen jeweils für eine bestimmte biochemische Reaktion. Dabei können diese, wie auch Metabolite, an mehreren Positionen der Karte auftreten (gleiche Metabolite sind durch weiße Linien verbunden). Mittels JavaScript lassen

sich die Namen und zugehörige Pathways in der Statuszeile des Browsers anzeigen, und per Mausklick gelangt man direkt zum gewünschten Pathway-Map.

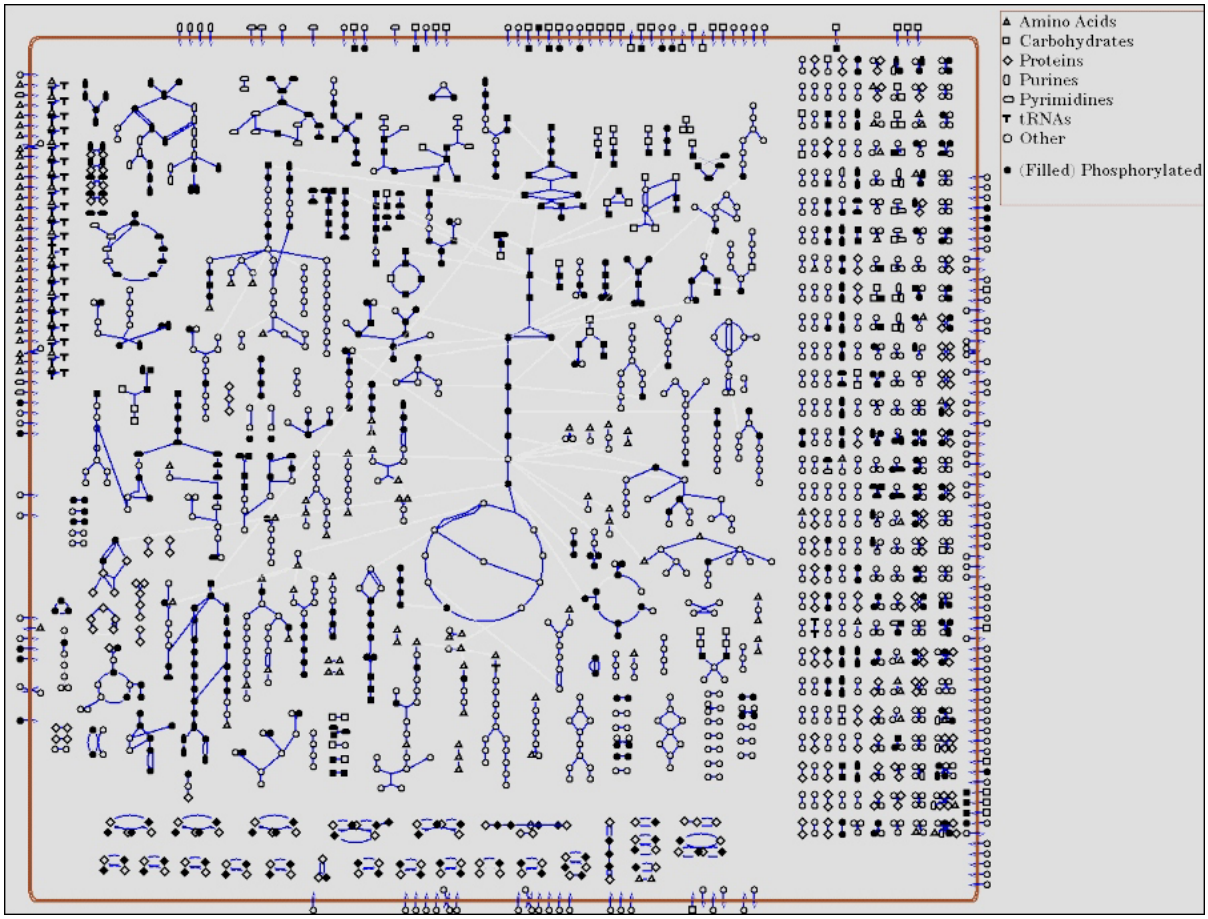


Abb. 13: Metabolic-Overview-Diagramm für E. coli, [EcoCyc] Glycolyse und Citratzyklus sind in der Mitte dargestellt, links davon Pfade der Biosysnthese, auf der rechten Seite die katabolischen Pfade. Im unteren Bereich sind Pfade zur Signaltransduktion angeordnet. Reaktionen, die bisher noch nicht zu Pathways zugeordnet werden konnten, sind am rechten Rande aufgeführt. In der durch den äußeren Rahmen symbolisierten Zellmembran sind einige Transportreaktionen gezeigt.

Eine interessante Anwendung des EcoCyc Systems ist es, importierte Expressionsdaten von E. coli innerhalb des Overview-Diagramms anzuzeigen (Pathway Tools Overview Expression Viewer). Dazu wird erhöhte oder verminderte Expression Enzym-kodierender Gene farbkodiert auf die entsprechenden Reaktionen im Diagramm übertragen. Dadurch kann unmittelbar untersucht werden, welche Pathways unter bestimmten experimentellen Bedingungen an- oder abgeschaltet sind. Auch zeitliche Veränderungen in der Expression können in Form von Animationen visualisiert werden. Beispiele hierzu sind unter http://biocyc.org/ov-expr.shtml zu finden.

Wie auch schon bei KEGG ist die EcoCyc-Datenbank zusätzlich im flat-file Format verfügbar und für die lokale Installation der Datenbank samt Pathway-Tools Software vorbereitet. Diese bietet weitergehende Abfrage- und Analysemethoden an, die über die auf der BioCyc-Webseite angebotenen Features hinausgehen.

3.4 Zukünftige Herausforderungen

Während biochemische Reaktionsnetze auf Stoffwechselebene schon seit langem untersucht werden, beschäftigt man sich mit den regulatorischen Netzen noch nicht so lange. Zukünftige Herausforderungen an Pathway-Datenbanken sind daher beispielsweise die Modellierung von den viel größeren Signalübertragungsnetzen eukaryotischer Organismen und die Entwicklung von Methoden zur Visualisierung und Navigation dieser großen Netzwerke [Kar01].

Ein wesentlicher Punkt für den Gewinn neuer Erkenntnisse ist auch die Unterstützung des Datenaustausches zwischen den sehr unterschiedlichen Datenbanken und Anwendungsprogrammen. Dies erfordert Informations-Standards wie z.B. SBML (Systems Biology Markup Language, [SBML]) und einheitliche Begriffssysteme (Ontologien).

Vom Entwurf neuer Algorithmen zur Pathway-Analyse, vor allem zur Unterstützung des Arzneimittel-Designs (u.a. Bestimmung optimaler Angriffspunkte (drug targets) in Netzen pathogener Organismen), verspricht man sich Behandlungserfolge bei bisher nur schlecht behandelbaren Krankheiten.

4. Zusammenfassung

In dieser Arbeit wurden Pathway-Datenbanken unter dem Schwerpunkt ihrer Anwendung als wichtiges Hilfsmittel der Bioinformatik und Systembiologie vorgestellt.

Pathway-Datenbanken verfolgen mehrere Zielsetzungen. Zum einen sind sie enzyklopädische Wissensbanken und Referenzen zu Pathway relevanten Informationen, die Wissenschaftlern den Zugang zu spezifischen Fakten und auch globalen Zusammenhängen von Reaktionsnetzen bieten. Zum anderen wird bei entsprechender interner Strukturierung und Repräsentation der Daten die rechnergestützte Analyse und Operation auf ihnen ermöglicht, die beispielsweise zur leichteren Interpretation von Expressionsdaten dienen kann.

Anhand von zwei wichtigen Vertretern, KEGG und EcoCyc, wurden Pathway-Daten und ihre Visualisierung und Repräsentation in Datenbanksystemen beschrieben. Unter dem Aspekt des benutzerfreundlichen Zugriffs und der integrierten Datenfülle ist KEGG besonders positiv hervorzuheben. EcoCyc dagegen bietet gerade durch die dynamische Erstellung der Pathway-Abbildungen mit Unterstützung verschiedener Detailansichten große Vorteile in der Navigation, wenn man sich erst einmal mit dem System vertraut gemacht hat und spezifisch an E.coli interessiert ist.

Literaturverzeichnis

[Ba03] Baxevanis, A.D.: The Molecular Biology Database Collection: 2003 update, *Nucleic Acids Research*, 31(1): 1-12, 2003, http://nar.oupjournals.org

[Ba00] Bahnsen, U.: Im Dickicht der Proteine. *Die Zeit*, Nr. 29, 14. Juli, Seiten 33-34, 2000

[BM02] Bundesministerium für Bildung und Forschung (Hg.): *Systembiologie. Systeme des Lebens*, Bonn, 2002

[GJ02] Gibas, C./Jambeck, P.: *Einführung in die Praktische Bioinformatik*, Köln, 2002

[Hu93] Hunter, L.: *Molecular Biology for Computer Scientists*, in: Artificial Intelligence and Molecular Biology, Cambridge, 1993, verfügbar unter: http://www.aaai.org/Library/Books/Hunter/hunter.html

[Kan01] Kanehisa, M: Knowledge-based Prediction of Cellular Functions from Genome Information, Frank and Bobbie Fenner Conference, Canberra, September 2001

[Kar01] Karp, P.D.: Pathway databases: A case study in computational symbolic theories, *Science*, 293: 2040-2044, 2001

[KG+02] Kanehisa, M./Goto, S./Kawashima, S./Nakaya, A.: The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30: 42-46, 2002

[KK+99] Karp, P.D./Krummenacker, M./Paley, S./Wagg, J.: Integrated pathway/genome databases and their role in drug discovery, *Trends in Biotechnology*, 17(7): 275-281, 1999

[KR+02] Karp, P.D./Riley, M./Saier, M./Paulsen, I.T./Paley, S./Pellegrini-Toole, A.: The Ecocyc Database, *Nucleic Acids Research*, 30(1): 56, 2002

[Mi99] Michal, G.: *Biochemical Pathways. Biochemie-Atlas*, Heidelberg, 1999.

[RH01] Rehm, H./Hammar, F.: *Biochemie light*, Frankfurt am Main, 2001

[Sc01] Schreiber, F.: Visualisierung biochemischer Reaktionsnetze, http://elib.ub.uni-passau.de/opus/volltexte/2001/21/index.html

[Sh02] Shrager, J.: Just Enough Molecular Biology for Computer Scientists, http://aracyc.stanford.edu/~jshrager/jeff/mbcs/

[VL+00] Vogelstein, B./Lane, D./Levine, A.J.: Surfing the p53 network, *Nature*, 408: 307-310, 2000

[WD01] Wittig, U./De Beuckelaer, A.: Analysis and comparison of metabolic pathway databases, *Briefings in Bioinformatics*, 2(2): 126-142, 2001

Verzeichnis verwendeter Web-URLs

[BMBPC] Boehringer Mannheim Biochemical Pathways Chart http://biochem.boehringer-mannheim.com/publications/metamap.htm

[BRENDA] BRENDA - The Comprehensive Enzyme Information System http://www.brenda.uni-koeln.de/

[BioCarta] BioCarta - Charting Pathways of Life, http://www.biocarta.com/genes/

[BioCyc] BioCyc Knowledge Library, http://biocyc.org/

[CSNDB] Cell Signaling Networks Database, http://geo.nihs.go.jp/csndb/

[DBGET] DBGET Integrated database retrieval system, http://www.genome.ad.jp/dbget/

[EcoCyc] EcoCyc: Encyclopedia of Escherichia coli Genes and Metabolism http://biocyc.org/ecocyc/

[ExPASy-ENZYME] http://www.expasy.org/enzyme/

[ExPASy-Pathways] http://www.expasy.org/cgi-bin/search-biochem-index

[GenomeNet] GenomeNet-Service, http://www.genome.ad.jp

[IUBMB] NC-IUBMB Enzyme Nomenclature, http://www.chem.qmul.ac.uk/iubmb/enzyme/

[IUBMB2] International Union Of Biochemistry And Molecular Biology, Recommendations on Biochemical & Organic Nomenclature, Symbols & Terminology etc., http://www.chem.qmw.ac.uk/iubmb/

[Kanehisa Laboratory] http://kanehisa.kuicr.kyoto-u.ac.jp/

[KEGG] KEGG: Kyoto Encyclopedia of Genes and Genomes http://www.genome.ad.jp/kegg/

[MALARIA] Malaria Parasite Metabolic Pathways, http://sites.huji.ac.il/malaria/

[MetaCyc] MetaCyc Metabolic Pathway Database, http://biocyc.org/metacyc/

[PTools] Pathway Tools Information Site, http://bioinformatics.ai.sri.com/ptools/

[SBML] The Systems Biology Markup Language (SBML), http://www.sbw-sbml.org/sbml/docs/index.html

[SPAD] Signaling Pathway Database, http://www.grt.kyushu-u.ac.jp/spad/

[WIT] WIT, http://wit.mcs.anl.gov/WIT2/

Auswahl aktueller Forschungsaktivitäten und Konferenzen zum Thema:

- The BioPathways Consortium, http://www.biopathways.org/
- PSB 2003: 8th Pacific Symposium on Biocomputing, Hawaii (January 3-7, 2003) http://psb.stanford.edu/
- International workshop on Computational Methods in Systems Biology, Rovereto, Italy (February 24-26, 2003), http://www.unitn.it/convegni/cmsb.htm
- RECOMB 2003: 7th Int. Conf. on Research in Computational Molecular Biology, Berlin (April 10-13, 2003), http://www.ctw-congress.de/recomb/
- ISMB 2003: 11th Int. Conf. on Intelligent Systems for Molecular Biology, Brisbane, Australia (June 29-July 3, 2003), http://www.iscb.org/ismb2003/
- 3rd ASM/TIGR Conference on Microbial Genomes, New Orleans (January 29-February 1, 2003), http://www.tigr.org/conf/mg/index.htm
- International Conference on Systems Biology 2004 http://www.icsb2004.org/