Hochschule für Angewandte Wissenschaften Hamburg

Fakultät Life Sciences

in Zusammenarbeit mit der Forschungsgruppe Biosystems Engineering
der Bogazici University Istanbul

Erstellung einer interaktiven Datenbank mit grafischer Benutzeroberfläche für Protein-Protein-Interaktionen zwischen Strukturen von pathogenen Erregern und vom Menschen

Bachelorarbeit

im Studiengang Biotechnologie

vorgelegt von

Ali Semih, Sayilirbas
Matrikelnummer: 1855218

Hamburg, 21. August 2013

Erstgutachter: Prof. Dr. Paul, Scherer (HAW Hamburg)
Zweitgutachter: Dr. Saliha, Durmus Tekir (Bogazici University Istanbul)

VORWORT

“Keine Schuld ist dringender als die, Dank zu sagen” (Marcus Tullius Cicero). Und keine Schuld begleiche ich mit mehr Freude als diese.

An erster Stelle möchte ich mich bei meinem Betreuer an der HAW Hamburg, Prof. Dr. Paul Scherer, für seine Unterstützung und seine wertvollen Anregungen zu dieser Arbeit bedanken. Meiner Betreuerin an der Bogazici University Istanbul, Dr. Saliha Durmus Tekir, danke ich herzlichst für die stets professionelle und doch entspannte Zusammenarbeit sowie für das Vertrauen, das sie mir während der ganzen Zeit entgegengebracht hat, sodass ich weitestgehend selbständig und meine eigenen Ideen einbringend arbeiten konnte. An dieser Stelle möchte ich ihr auch zur Fertigstellung ihrer Doktorarbeit gratulieren, in deren Rahmen die vorliegende Arbeit überhaupt erst ermöglicht wurde.

Meinen Eltern Elif und Izzettin Sayilirbas sowie meiner Schwester Yasemin Sayilirbas bin ich für ihre emotionale Unterstützung und ihren Glauben an mich und zu größtem Dank verpflichtet.

Nicht zuletzt danke ich meiner Lebensgefährtin Timea Markus für ihre liebevolle und geduldige Unterstützung. Ihre Anwesenheit hat stressige Zeiten erleichtert und mir Motivation gegeben.

INHALTSVERZEICHNIS

ABKÜRZUNGEN & AKRONYME

CSS	Cascading Style Sheets
DB	Datenbank
DBMS	Datenbankmanagementsystem
DBS	Datenbanksystem
DDL	Data Definition Language
ERM	Entity-Relationship-Modell
GBO	Grafische Benutzeroberfläche
HIV/HI-Virus	Humanes Immundefizienz-Virus
HTML	Hypertext Markup Language
PHP	Hypertext Preprocessor
PPI	Protein-Protein-Interaktion
SQL	Structured Query Language
UML	Unified Modeling Language
XML	Extensible Markup Language

FACHBEGRIFFE

Attribut	Merkmal eines Objekts/einer Entität
Back-End	Bei Datenbankanwendungen die Datenbank
Bottleneck	Proteine von geringer Konnektivität, die brückenartig Unternetzwerke eines Gesamtnetzwerkes verknüpfen
Constraint/Zwangsbedingung	Bedingung zu einer Tabellenspalte, die zwingend von den Werten dieser Spalte erfüllt werden muss
Curated (database)	Datenbank, deren Inhalt zuvor von Fachpersonal geprüft und aufbereitet wird
Data Definition Language	Computersprache zur Beschreibung von Datenstrukturen
Datenbanksystem	System zu elektronischen Datenverwaltung, welches den Datenbestand und das Datenbankmanagementsystem umfasst
Datenbankmanagementsystem	Software zur Strukturierung und Verwaltung einer Datenbank
Datenmodell	Modell von zu beschreibenden und verarbeitenden Daten eines realen Anwendungsbereiches
Entität	In der Datenmodellierung ein Objekt
Entity-Relationship-Modell	In Diagrammform dargestelltes Modell zur Beschreibung eines Datenmodells
Evaluierung	Systematische Untersuchung und Auswertung
Foreign Key/Fremdschlüssel	Attribut, dass in relationalen Datenbanken auf einen Primärschlüssel verweist
Front-End	Bei Datenbankanwendungen die grafische Benutzeroberfläche
Hub	Stark vernetztes Protein innerhalb eines Proteinnetzwerkes
Implementierung	Umsetzung eines Softwareentwurfs; Programmierung
Interaktom	Gesamtnetzwerk der Protein-Protein-Interaktionen
Interspezifisch	Zwischenartlich; zwischen verschiedenen Arten
Intraspezifisch	Innerartlich; innerhalb ein und derselben Art
Kardinalität	Angaben, wie Entitäten mengenmäßig miteinander zusammenhängen
Kurator (Datenbank)	Fachperson, die Daten prüft und aufbereitet
Martin-Notation	Notationsart zur Darstellung von Kardinalitäten in einem Entity-Relationship-Modell

Normalisierung	Aufteilung von Attributen/Tabellenspalten in mehrere Entitätstypen/Tabellen gemäß bestimmter Regeln, sodass vermeidbare Redundanzen beseitigt werden
Primary Key/Primärschlüssel	Attribut, dass in relationalen Datenbanken zu eindeutigen Identifizierung von Datensätzen dient
Relational	Auf Datenbanktabellen und Beziehungen zwischen den Tabellen beruhend
Schnittstelle	Hier Teil der Software, welche die Kommunikation mit der Datenbank ermöglicht
Sicht	Logische Relation aus Datenbanktabellen, die über eine Abfrage im Datenbankmanagementsystem gespeichert wird; virtuelle Tabelle
Target	Zielmolekül, an das sich ein pharmakologischer Wirkstoff binden kann
Text Mining	Algorithmus-basierte Suche und Identifikation von Textstrukturen

1. EINLEITUNG

Interspezifische Protein-Protein-Interaktionen (PPI) zwischen Proteinen von Pathogenen und humanen Proteinen spielen eine entscheidende Rolle bei Infektionskrankheiten, denn das Eindringen pathogener Erreger in ihre Wirtsorganismen und das Verharren in diesen wird oft erst durch solche PPI ermöglicht. Während die meisten PPI-Studien auf die Aufklärung spezieller Infektionsmechanismen wie z. B. das Eindringen von Viren in menschliche Wirtszellen durch Membran-assoziierte Interaktionen mit Rezeptorproteinen ausgerichtet sind, wird seit einigen Jahren in großangelegten Studien die systematische Identifikation aller Interaktionen zwischen Proteinen pathogenen Ursprungs und humanen Proteinen angestrebt. Auf diese Weise experimentell nachgewiesene PPI bilden die Grundlage für die Modellierung von interspezifischen Interaktionsnetzwerken, deren Kartierung und Analyse einen systembiologischen Einblick in die komplexen Infektionsmechanismen ermöglichen und zu neuen Erkenntnissen führen, wie Infektionskrankheiten verhindert oder behandelt werden können, um ganz neue Pharmazeutika der nächsten Generation zu entwickeln.

Die Fortschritte in den Hochdurchsatz-Technologien zum Nachweis von molekularen Interaktionen führen seit einigen Jahren kontinuierlich zu immer größeren Mengen an PPI-Daten, die für weiterführende bioinformatische und pharmakologische Forschungen öffentlich und umfassend zur Verfügung gestellt werden müssen. Zwar existieren bereits einige wenige öffentliche Datenbankanwendungen für Interaktionen zwischen Pathogenen und dem menschlichen Organismus, doch diese sind allesamt entweder nicht umfangreich genug, nicht aktuell oder bieten nicht genügend Möglichkeiten zur Datenrecherche und -analyse. Da interspezifische PPI-Daten über diese Datenbanken verstreut sind und nicht einheitlich verwaltet werden, sind Nachforschungen in diesem Gebiet noch sehr mühsam. Aus diesem Grund wurde in dieser Bachelor-Arbeit die neue Software PHISTO (Pathogen Host Interaction Search Tool) als neue Datenbankanwendung entwickelt, welche den Anwendern Zugang zu der umfangreichsten Sammlung von konsistenten und aktuellen PPI zwischen Viren, Bakterien, Pilzen, Protozoen und dem menschlichen Organismus bietet. Die Grundlage dafür bildet eine im Rahmen dieser Arbeit neu entwickelte, sekundäre PPI-Datenbank, deren Inhalt auf vereinheitlichten PPI-Daten aus anderen öffentlichen Datenbanken beruht. Der Zugang zu dieser Datenbank wird über die Webseite www.phisto.org ermöglicht. Diese bietet den Anwendern eine übersichtliche grafische Benutzeroberfläche mit verschiedenen Optionen zur einfachen sowie schnellen Datenrecherche. Darüber hinaus stellt PHISTO eine Plattform mit einer erweiterbaren Archi-

tektur dar und kann somit einfach sowohl um weitere Daten als auch um weitere Funktionalitäten zur Analyse und Darstellung der Daten ergänzt werden. Der Autor dieser Bachelor-Arbeit ist dementsprechend im Impressum der Webseite aufgeführt.

2. THEORETISCHE GRUNDLAGEN

2.1. Protein-Protein-Interaktionen (PPI) zwischen pathogenen Strukturen und dem Menschen

Protein-Protein-Interaktionen (PPI) sind biochemische und physikalische Wechselwirkungen zwischen zwei Proteinen. Sie spielen bei vielen zellulären Stoffwechselprozessen aller Organismen eine Schlüsselrolle, unter anderem bei der Signaltransduktion, der Genexpression, der Enzymregulation und bei immunologischen Prozessen (Domling et al, 2013, S. 2). Darüber hinaus sind interspezifische PPI, also Interaktionen zwischen zwei Proteinen, die von unterschiedlichen Spezies stammen, bei Infektionsmechanismen pathogener Organismen von großer Bedeutung. Im Folgenden soll der Fokus auf solchen Interaktionen liegen.

Infektionsmechanismen von pathogenen Mikroorganismen beruhen auf verschiedensten molekularen Interaktionen zwischen den Proteinen, Polysacchariden, Nukleinsäuren, Lipiden und komplexen Liganden des Erregers und denen des Wirts. Von diesen gelten PPI als die wichtigsten und daher bisher auch am meisten erforschten Interaktionen (Stebbins, 2005). Ein Beispiel, welches die Bedeutung von interspezifischen PPI für Infektionsmechanismen verdeutlicht, ist das Eindringen von Salmonellen in menschliche Darmzellen (Abb. 1, orange umkreister Bereich). Der Invasionsmechanismus von Salmonellen, welcher der spezifischen Anheftung des Bakteriums an polysaccharidische Oberflächenantige folgt, zielt auf das Aktin-Zellskelett von Darmzellen. Aktin ist ein Protein, das feine und dynamische Faserstrukturen bildet, welche der Zelle Halt geben und sie gleichzeitig beweglich machen. Salmonellen synthetisieren bei Zellkontakt die Zellinvasion-Proteine SipA und SipC, welche auf der Zelloberfläche mit Aktin-Proteinen interagieren: SipC bewirkt die Keimbildung aus Aktin-Monomeren und die Polymerisation, SipA bindet Aktinfasern und moduliert deren Bündelung. Nach Verankerung in das Zytoskellet bildet SipC ein Sekretionssystem für die Einschleusung Virulenz-assoziierter Proteine wie z.B. SigD. Die Interaktionen von SipA und SipC mit Aktin-Proteinen des Zytoskellets führen zu örtlichen Ausstülpungen der Zellmembran. Die Bakterien lassen sich von diesen Ausstülpungen umschließen und so ins Zellinnere aufnehmen (Zhang et al., 2008).

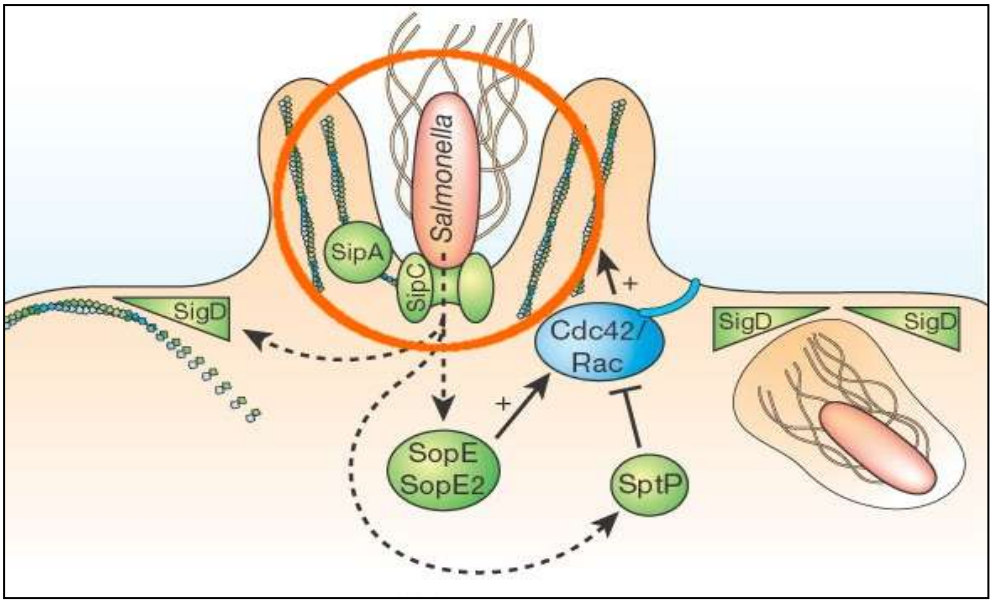


Abb.1 **Beispiel für eine Protein-Protein-Interaktion: Eindringen einer Salmonellenzelle in eine Wirtszelle**

Nach der Anheftung des Bakteriums an die Darmzelloberfläche über antigene Gruppen ermöglichen Protein-Protein-Interaktionen (PPI) auf und in der Witszelle das Eindringen in die Wirtszelle. SipA und SipC, von Salmonellen synthetisierte Zellinvasion-Proteine, interagieren mit Aktin-Proteinen des Zytoskellets (orange umkreist), indem sie diese binden und dadurch örtliche Membranausstülpungen bewirken. SigD löst die Verbindung zwischen der Membran und dem Aktinskelett. Cdc42/Rac sind Rho-GTPasen, die in der Zelle als Regulaturen der Singaltransduktion für die Organisation und den Umbau des Cytoskeletts dienen. Diese werden durch SopE und SopE2 aktiviert, sodass sie den Umbau des Aktinskeletts beschleunigen. SptP inaktiviert CDC42/Rac. Das Zusammenspiel dieser PPI führt letztlich dazu, dass die Salmonellenzelle von den Ausstülpungen umschlossen und in die Darmzelle aufgenommen wird.

Zhang et al., *Salmonella Typhi: from Human Pathogen to a Vaccine Vector*, 2008

Um PPI nachzuweisen, werden verschiedene experimentelle Methoden verwendet, von denen die gängigsten das Hefe-Zwei-Hybrid-System, die Affinitätsreinigung mit anschließender Massenspektroskopie, die Co-Immunoprezipitation und die Röntgenkristallographie sind (Giralt et al. ,2011, S. 4-13). Bei der Entdeckung einer PPI werden die experimentellen und molekularbiologischen Daten, auf die im Verlauf dieser Arbeit noch genauer eingegangen wird, als Nachweis veröffentlicht. Dabei sind die Daten einer experimentell nachgewiesenen PPI nicht als genaues Abbild der biologischen Interaktion zu verstehen, da sich Proteine unter physiologischen Bedingungen anders verhalten als unter experimentellen Bedingungen, sondern vielmehr als artifizielle Interaktion, die mit einer bestimmten Sicherheit auch unter physiologischen Bedindungen auftritt. Daher werden Interkationen oft mehrfach mit unterschiedlichen Methoden untersucht.

Während die Daten einzelner interspezifischer PPI wie im obigen Beispiel für die Aufklärung einzelner Infektionsmechanismen betrachtet werden, können sie in ihrer Gesamtheit verwendet werden, um das zwischenartliche Interaktom (das Netzwerk aller Interaktionen) darzustellen, welches als molekularbiologisches Modell für bioinformatische Analysen und die Identifikation neuer Targets (Zielstrukturen, an die sich ein pharmakologisch aktiver Wirkstoff binden kann) dienen kann (s. Abb. 2). Die noch sehr junge Interaktomforschung hat bereits einige Erfolge wie die Identifikation von humanen Proteinen zu verzeichnen, die im Interaktom einen stark vernetzten Knoten oder einen Engpass darstellen und häufig das Ziel von Infektionsmechanismen verschiedener Pathogene (z. B. von HI- oder Herpes-Viren) sind (Durmus Tekir S., 2013, S. 21). Solche Erkenntnisse können für alternative Ansätze zur Identifikation neuer, potentieller Targets genutzt werden. Beispielsweise könnten humane Proteine in Frage kommen, die für das Verharren und die Vermehrung der Erreger verantwortlich sind. Falls solch ein Protein für die wichtigsten Zellfunktionen entbehrlich ist, könnte es inaktiviert werden, um den Erreger unschädlich zu machen (Bünninge, 2011). Diese neuartige Herangehensweise hat gegenüber der konventionellen Methode einen erheblichen Vorteil, denn bisher werden Pathogene bekämpft, indem deren essentielle Zellstrukturen anvisiert werden, doch dadurch wird ein Selektionsdruck auf die Pathogene verursacht, der immer resistentere Stämme hervorbringt (insbesondere bei RNA-Viren mit hohen Mutationsraten). Humane Proteine als Target zu verwenden, würde nicht nur die Bekämpfung resistenter Stämme ermöglichen, sondern auch die Entstehung dieser eindämmen (Bünninge, 2011).

Zwischenartliche PPI sind der Wissenschaft durch Forschungen über Infektionsmechanismen einzelner Pathogenstrukturen schon lange bekannt, doch der Trend zu großangelegten Studien zur Aufklärung des gesamten Interaktoms von Pathogenen ist recht neu, denn erst mit den Fortschritten in der Laborautomatisierung und den Entwicklungen von Hochdurchsatz-Technologien bot sich der Forschung eine Möglichkeit, in kurzer Zeit tausende biochemische und molekularbiologische Tests zum Nachweis von PPI durchzuführen. Bisher lag der Fokus solcher großen Studien auf einer überschaubaren Anzahl bakterieller und viraler Erregern, sodass die Interaktome und damit die Infektionsmechanismen der meisten Pathogene noch größtenteils unaufgeklärt oder gänzlich unbekannt bleiben. Allerdings deuten die in den letzten 10 Jahren steigenden Zahlen an wissenschaftlichen Veröffentlichungen über PPI zwischen Pathogenen und dem menschlichen Organismus darauf hin, dass in Zukunft – nicht zuletzt durch das steigende Interesse der Pharmaforschung – noch weitaus intensiver in diesem Gebiet geforscht werden wird und dass mit der Entdeckung weiterer experimentell erwiesener

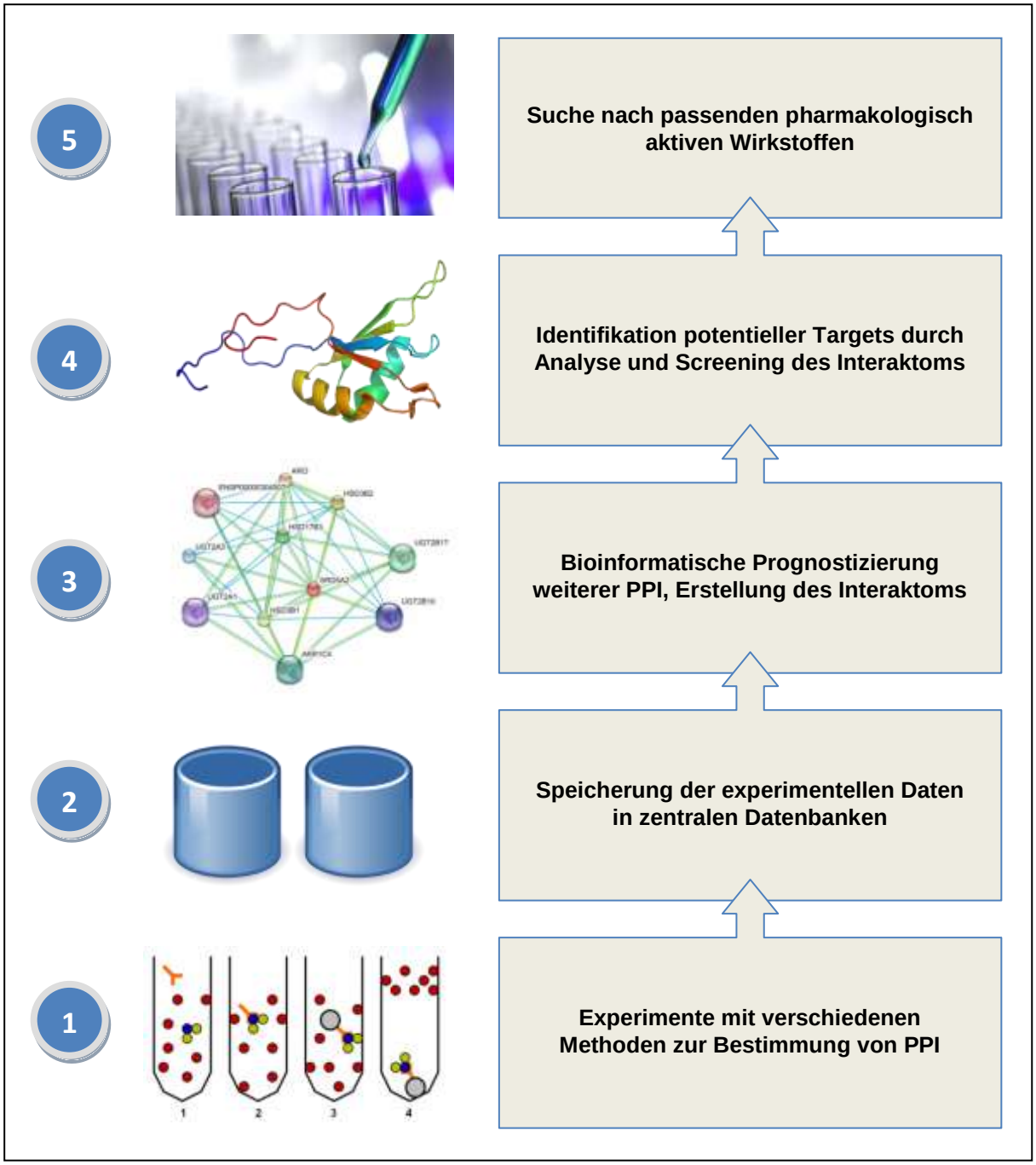


Abb. 2 Von der Protein-Protein-Interaktion zum pharmazeutischen Target

Interspezifische PPI zwischen Pathogenen und dem menschlichen Organismus könnten in Zukunft für die pharmazeutische Biotechnologie bei der Identifikation neuer, potentieller Targets für Wirkstoffe (drug targeting) von großer Bedeutung sein. Die wesentlichen Schritte zur Identifikation neuer Targets mit Hilfe von PPI sehen wie folgt aus:

1. Mit verschiedenen experimentellen Methoden wie z. B. dem Hefe-Zwei-Hybrid-System werden PPI nachgewiesen und charakterisiert.
2. Die Daten experimentell nachgewiesener Interaktionen werden in zentralen Datenbanken gespeichert und gepflegt.
3. Die Daten dienen als Grundlage für bioinformatische Methoden zur Prognostizierung weiterer PPI und zur Erstellung des Interaktoms.
4. Das Interaktom wird analysiert, um potentielle Targets für pharmakologisch aktive Wirkstoffe zu finden.
5. Die Wirkung verschiedene Substanzen werden mit dem Target getestet, um aus ihnen einen passenden pharmakologischen Wirkstoff zu identifizieren.

PPI zu rechnen ist (Durmus Tekir S., 2013, S. 21). Ein weiterer Ansatz zur Ermittlung von PPI ist der computergestützte Vergleich der Strukturen von Proteinen, bei denen eine Interaktion nachgewiesen werden konnte, mit denen ähnlicher, bisher nicht experimentell untersuchter Proteine (Nussinov et al., 2009, S. viii). Findet sich z. B. sowohl in dem interagierenden Protein als auch in dem untersuchten Protein eine sehr ähnliche Domäne, kann prognostiziert werden, dass mit einer gewissen Wahrscheinlichkeit auch das untersuchte Protein wie das Vergleichsprotein Interaktionen aufweisen wird (Szklarczyk et al., 2011).

Sowohl experimentell nachgewiesene als auch prognostizierte PPI werden in Zukunft zu noch größeren Datenmengen führen. Um die Daten analysieren und vergleichen zu können, müssen diese informationstechnologisch vereinheitlicht, gespeichert und verfügbar gemacht werden (Srivastava, 2005, S.164). Da aber bisher keine weltweit zentrale Datenbank existiert, bei der sämtliche PPI sofort nach ihrer Entdeckung berichtet, geprüft und gespeichert wurden, bleiben zunächst viele in der Vergangenheit nachgewiesene PPI in der wissenschaftlichen Literatur verborgen. Diese mittels automatisierter Computerprogramme zu suchen und zu finden (Text Mining) ist von großer Bedeutung und bisher noch eine große Herausforderung (Cannataro & Guzzi, 2011, S. 10).

Es gibt bereits einige online verfügbare PPI-Datenbanken und Anwendungen, die entweder als primäre Datenbanken ihre Daten aus wissenschaftlichen Arbeiten der Interaktomforschung beziehen und von eigenen Kuratoren gepflegt werden oder als sekundäre Datenbanken von anderen, hauptsächlich primären Datenbanken PPI-Daten sammeln. Zu den primären PPI-Datenbanken gehören MINT (Licata et al., 2012), DIP (Salwinski et al., 2004), IntAct (Kerrien et al., 2012), BIND (Alfarano et al., 2005), BioGrid (Stark et al., 2011), Reactome (Croft et al., 2011), STRING (Szklarczyk et al., 2011) und PATRIC (Gillespi et al., 2011). Zu den sekundären Datenbanken gehören iRefindex (Razick et al., 2008) APID (Prieto & Rivas, 2006), MPIDB (Goll et al., 2008), HPIDB (Kumar & Nanduri, 2010). Einer Übersicht dieser Datenbankanwendungen samt ihrer für einen Vergleich relevanten Eigenschaften ist in Tab. 1a und Tab. 1b zu sehen.

Datenbank	Daten	Organismen	Datenquelle	Aktualität	Datenrecherche	Darstellung	Zusätzliche Funktionen
MINT	Exp. PPI	Mensch, Säugetiere, Viren, Modellorganismen	Curated	unbekannt	Suche nach verschiedenen Kriterien	Liste von P. mit Verweisen zu interagierenden P.	Download aller Daten, BLAST, Verweise auf andere wissenschaftliche DB, Netzwerkvisualisierung
DIP	Exp. PPI	Mensch, Säugetiere, Modellorganismen	Curated, data mining	Feb. 2013	Suche nach verschiedenen Kriterien	Liste von P. mit Verweisen zu interagierenden P.	Download aller Daten, BLAST, Verweise auf andere wissenschaftliche DB
IntAct	Exp. MI	Verschiedene	Curated	regelmäßig	Suche nach verschiedenen Kriterien, Browse	Tabelle von P. mit interagierenden P.	Download aller o. recherchierter Daten, Verweise auf andere wissenschaftliche DB
BIND	Exp. MI	Verschiedene	Curated	offline	-	-	-
BioGrid	Exp. MI	Verschiedene	Curated	regelmäßig	Suche nach verschiedenen Kriterien, Browse	Liste von P. mit Verweisen zu interagierenden P.	Download aller o. recherchierter Daten, Verweise auf andere wissenschaftliche DB, Netzwerkvisualisierung
Reactome	Exp. MI	Mensch, Säugetiere, Pathogene	Curated	regelmäßig	Suche nach verschiedenen Kriterien, Browse	Interaktions-Diagramme	Download aller Daten, Verweise auf andere wissenschaftliche DB
STRING	Exp. PPI u. Com. PPI	Verschiedene	Curated	Jan. 2013	Suche nach verschiedenen Kriterien	PPI-Netzwerk	Download aller o. recherchierter Daten, Verweise auf andere wissenschaftliche DB, Prognose von PPIs
PATRIC	Exp u. Com. PPI	Bakterien, Viren	Curated	regelmäßig	Suche nach verschiedenen Kriterien, Browse	Tabelle von P. mit interagierenden P.	Download aller o. recherchierter Daten, Verweise auf andere wissenschaftliche DB, Netzwerkvisualisierung

Tab. 1a **Übersicht der für diese Arbeit betrachteten primären PPI-Datenbanken**

Datenbank	Daten	Organismen	Datenquelle	Aktualität	Datenrecherche	Darstellung	Zusätzliche Funktionen
iRefindex	Exp. PPI	Verschiedene	BIND, BioGrid, CORUM, DIP, HPRD, InateDB, IntAct, MatrixDB, MINT, Mpact, MPIDB, MPPI, OPHID	regelmäßig	-	-	Download aller Daten
APID	Exp. PPI	Verschiedene	BIND, BioGrid, DIP, HPRD, IntAct, MINT	2009	Suche nach verschiedenen Kriterien	Tabelle von P. mit interagierenden P.	Download recherchierter Daten, Netzwerkvisualisierung
MPIDB	Exp. PPI	Verschiedene	IntAct, DIP, BIND, MINT	2009	Suche nach verschiedenen Kriterien	Tabelle von P. mit interagierenden P.	Download recherchierter Daten
HPIDB	Exp. PPI	Pathogene	Reactome, BIND, MINT, BioGrid, VirHostNet, GENERIF, INTACT, PATRIC	Apr. 2013	Suche nach verschiedenen Kriterien	Tabelle von P. mit interagierenden P.	Download recherchierter Daten

Tab. 1b Übersicht der für diese Arbeit betrachteten sekundären PPI-Datenbanken

In jeder Zeile ist eine Datenbank mit den für diese Arbeit relevantesten Eigenschaften gelistet.

Bedeutung der Spalten

Datenbank: Name der betrachteten Datenbank; **Daten**: Art der gespeicherten Interaktionsdaten; **Organismen**: Organismen, zu denen Interaktionsdaten geführt werden; **Datenquelle**: Herkunft der gespeicherten Daten: Entweder von eigenen Kuratoren geprüfte und verwaltete Daten (engl.: curated), die aus wissenschaftlicher Literatur bezogen werden oder eigens durch computergestützte Analysemethoden ermittelt werden, oder von anderen Datenbanken bezogene Daten; **Aktualität**: Letztes Datum bzw. Frequenz der Aktualisierun; **Datenrecherche**: Möglichkeiten der Datenrecherche („Browse“ ist die englische und gängige Bezeichnung für das Betrachten der Daten anhand eines Kriteriums ohne einen vom Nutzer vorgegebenen Suchausdruck); **Darstellung**: Art der Darstellung recherchierter Daten; **Zusätzliche Funktionen**: Von der Datenbankanwendung zusätzlich zur Recherche zur Verfügung gestellte Funktionalitäten, die auf die Interaktionsdaten angewendet werden können

Abkürzungen

Com.: Durch computergestützte Analysen bestimmt; **DB**: Datenbank, **Exp**.: Experimentell bestimmt; **MI**: Molekulare Interaktionen aller Art; **P**.: Protein; **PPI**: Protein-Protein-Interaktionen

Keine dieser Datenbanken bietet momentan umfassend aktuelle Daten zu allen bisher bekannten humanpathogenen PPI und gleichzeitig ausreichende Funktionalitäten zur Datenrecherche und -analyse. Die primären Datenbanken sind sicherlich als zuverlässigste Quelle für PPI anzusehen, doch diese konzentrieren sich nicht auf interspezifische, sondern eher auf intraspezifische Interaktionen. Einzig PATRIC pflegt ausschließlich interspezifische PPI-

Daten, zu denen auch die in dieser Arbeit betrachteten PPI zwischen Pathogenen und dem menschlichen Organismus gehören, doch diese beschränken sich auf Bakterien und Viren. Ferner besitzen alle primären Datenbanken eigene Kuratoren, welche die in Frage kommenden Daten vor der Aufnahme in die Datenbank überprüfen, sodass sich auch schon aufgrund dieser Tatsache die gespeicherten Daten unterscheiden können. Eigens manuell durchgeführte Datenvergleiche ergeben, dass sich die Menge an PPI-Daten sogar bei der Betrachtung ein und desselben Organismus unterscheiden, was vermuten lässt, dass die Kuratoren unterschiedlich arbeiten und/oder unterschiedliche Quellen auswerten. Von den sekundären Datenbanken, die den Ansatz verfolgen, PPI-Daten aus primären Datenbanken zu vereinheitlichen und zentral zur Verfügung zu stellen, sind nur iRefIndex und HPIDB aktuell. Von diesen ist wiederum nur HPIDB als umfangreiche und zentrale Sekundärdatenbank für PPI zwischen Pathogenen und dem menschlichen Organismus anzusehen, doch die Anwendung ist recht simpel gehalten, bietet nicht genügend Rechercheoptionen und auch keine anderen bzw. erweiterten Möglichkeiten der Datenanalyse und -visualisierung. Aus diesem Grund wird mit dieser Arbeit die Software PHISTO geschaffen, welche als Recherche-optimierte und erweiterbare Datenbankanwendung für die aktuellsten und umfangreichsten PPI zwischen Pathogenen und dem menschlichen Organismus dienen soll.

2.2. Entwicklung von Datenbankanwendungen

Die Entwicklung von Datenbankanwendungen kann grundsätzlich in zwei Bereiche unterteilt werden: die Entwicklung der Datenbank und die Entwicklung des Anwendungsprogrammes in Form einer grafischen Benutzeroberfläche (GBO) (s. Abb. 3). Die Datenbank, meistens eine relationale Datenbank, welche als Sammlung von Tabellen und Beziehungen zwischen den Tabellen zu verstehen ist, dient zur effizienten und dauerhaften Speicherung sowie Verwaltung großer Datenmengen. Das Anwendungsprogramm ermöglicht den Zugriff und die bedarfsgerechte Darstellung der Daten. Beide Bereiche sind vier Abschnitte unterteilt: die Konzeption, die Implementierung, die Test- und Evaluierungsphase und die Wartungs- und Optimierungsphase. Die beiden aufeinanderfolgenden Schritte der Konzeption sowie der Implementierung bilden die eigentliche Entwicklungsarbeit und sollen daher am ausführlichsten von allen vier Abschnitten erläutert werden.

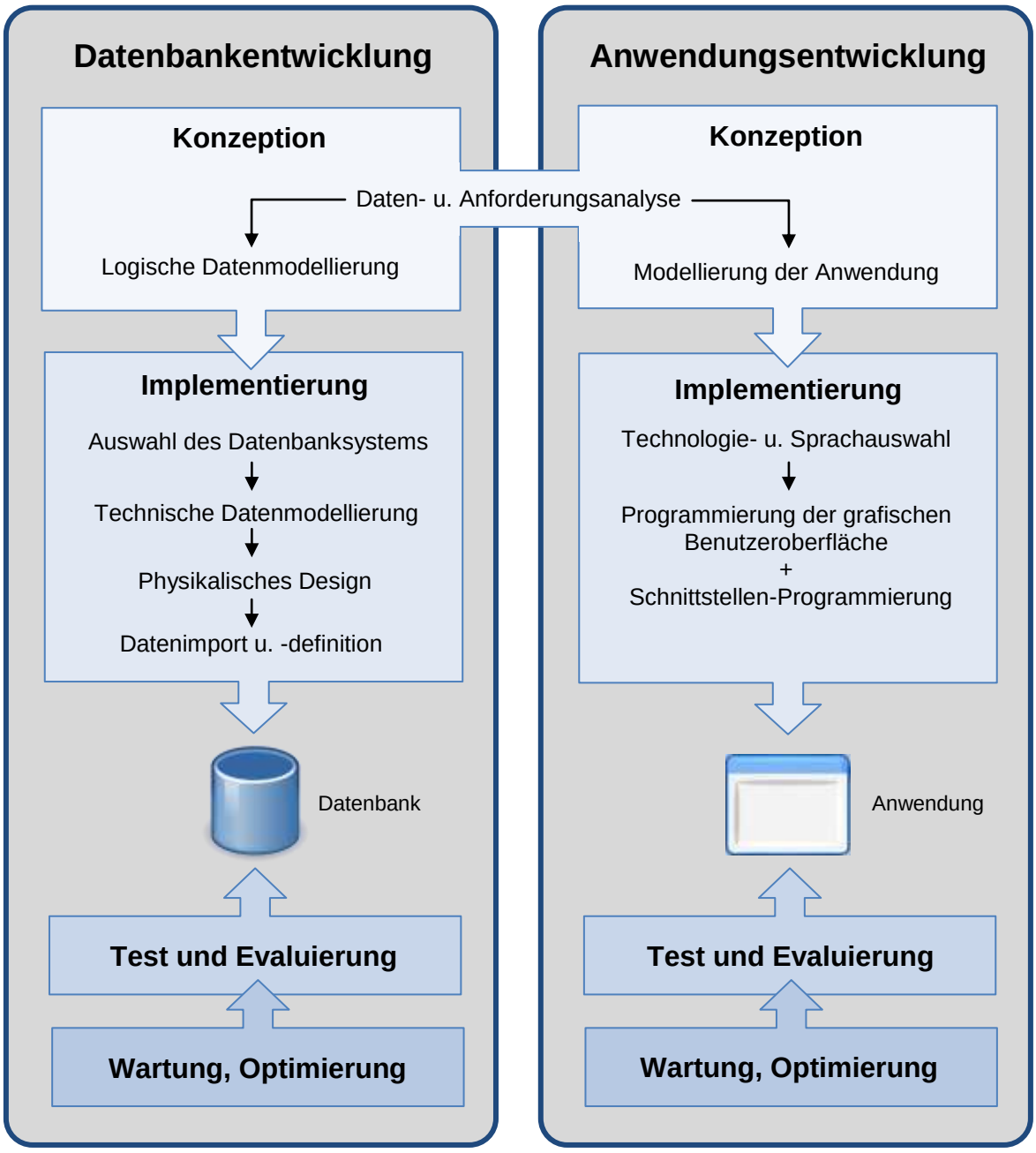


Abb. 3 Allgemeine Vorgehensweise bei der Datenbank- und Anwendungsentwicklung für Datenbankanwendungen

Die Entwicklung von Datenbankanwendungen kann grundsätzlich in zwei Bereiche unterteilt werden: die Entwicklung der Datenbank und die Entwicklung der Anwendung, welche typischerweise als grafische Benutzeroberfläche (GBO) erstellt wird. Beide Bereiche sind in vier Abschnitte unterteilt: die Konzeption, die Implementierung, die Test- und Evaluierungsphase und die Wartungs- und Optimierungsphase. Tests und Evaluierungen werden während der Entwicklung durchgeführt, während sich Wartungsarbeiten und Optimierungen über den gesamten Anwendungszeitraum erstrecken. Die Konzeptionsphase läuft in beiden Bereichen in zwei Teilschritten ab: Zuerst werden die Anforderungen und die gegebenen Daten analysiert. Dieser Schritt findet bereichsübergreifend statt, da die Anforderungen an die Anwendung maßgeblich die Anforderungen an die Datenbank bestimmen. Anschließend werden die Modelle für die Datenbank und die GBO erstellt. Die Implementierung beim Datenbankentwurf beginnt mit der Auswahl eines geeignete Datenbanksystems (DBS), in dem das logische Datenmodell in ein technisches Datenmodell übertragen wird (s. Abb. 4). Es folgt das physikalische Design des Datenschemas. Zuletzt können

gewünschte Daten mit Hilfe des Datenbankmanagementsystems importiert und definiert werden. Die Implementierung der Anwendung beginnt mit der Auswahl der Technologien und Sprachen zur Umsetzung des Anwendungsmodells. Anschließend werden die GBO und die Schnittstellen zur Datenbank programmiert.

2.2.1. Konzeption von Datenbankanwendungen

Am Anfang der Datenbankentwicklung steht die Erstellung eines konzeptionellen Datenmodells, das die realen Objekte und deren eventuelle Beziehungen zueinander möglichst genau beschreibt (Geisler, 2009, S. 298). Die eigentliche Realisierung der Datenbank ist hierbei zunächst nicht von Bedeutung, d.h. die Auswahl des Datenbankmanagementsystems spielt keine Rolle (Geisler, 2009, S. 298). Dadurch ist das konzipierte Modell plattformunabhängig und kann als Grundlage für die Realisierung in einem beliebigen System dienen. Notwendige Bedingung für ein gutes konzeptionelles Datenmodell ist die Umsetzung aller Anforderungen, die für die angestrebte Anwendung erforderlich sind (Geisler, 2009, S. 298). Dazu ist eine vorausgehende Daten- und Anforderungsanalyse nötig, wobei einerseits festgelegt wird, was überhaupt benötigt wird, um die Anwendung datenseitig zu realisieren, und andererseits analysiert wird, welche anwenderseitigen Operationen zukünftig auf die Daten realisiert werden sollen (Geisler, 2009, S. 298). Diese Herangehensweise, die Datenbankanwendung von Beginn an sowohl aus Datensicht (Back-End) als auch aus Anwendersicht (Front-End) zu betrachten, ermöglicht es, Anforderungen an die Anwendung sofort in das Datenmodell einfließen zu lassen. Wichtige Fragestellungen sind hierbei, ob Daten durch Anwender in der Datenbank gespeichert werden sollen, ob anwenderseitige, dynamische Recherchen in der Datenbank durchgeführt werden sollen und wie die Daten präsentiert werden sollen.

Um aus den erforderlichen Informationen, die anhand der Daten- und Anforderungsanalyse festgelegt wurden, ein relationales Datenmodell zu erstellen, wird das Konzept des Entity-Relationship-Modell (ERM) verwendet (Geisler, 2009, S. 301) . Für die Darstellung des Modells wird die grafische Modellierungssprache UML (Unified Modeling Language) verwendet, die als de-facto Standard für den objektorientierten Softwareentwurf gilt und als Ergebnis ein Diagramm des ERM liefert. Zuerst müssen alle eindeutig bestimmbaren Objekte gefunden werden. Ein Objekt entspricht dann im ERM einer Entität (Entity) (Meier, 2004, S. 16). Als nächstes müssen die für die Anwendung relevanten Eigenschaften der Objekte bestimmt und abstrahiert werden. Eine Eigenschaft entspricht dann im ERM einem Attribut (Geisler, 2009, S. 96). Nach Bestimmung der Attribute kann eine erste Normalisierung durchgeführt werden,

d.h. dass die Attribute wenn möglich so weit aufgeteilt werden, dass man eine Form erhält, die keine vermeidbaren Redundanzen mehr enthält (Kleinschmidt et al., 2004, S. 75 - 81). Zuletzt können Beziehungen zwischen den Entitäten bestimmt werden. Für die Festlegung dieser Beziehungen (Relationships) im ERM muss zunächst in jeder Entität ein Attribut als Primärschlüssel (Primary Key) vordefiniert werden (Meier, 2004, S. 16). Dieser dient zur eindeutigen Identifikation der Datensätze in dieser Entität. Anschließend muss in denjenigen Entitäten, zu denen eine Beziehung besteht, das Attribut vom selben Typ als Fremdschlüssel (Foreign Key) vordefiniert werden, sodass ein Verweis auf die Entität mit dem Primärschlüssel hergestellt wird. Der Verweis von Sekundär- auf Primärschlüssel ist einer der wichtigsten Maßnahmen zur Konsistenz-Erhaltung in relationalen Datenbanken und setzt immer voraus, dass ein Attributwert bereits in der Tabelle existiert, in der das Attribut als Primärschlüssel definiert ist, bevor dieser in anderen Tabellen als Wert eines Sekundärschlüssels gesetzt werden kann (Meier, 2004, S. 17). Dabei ist es hilfreich, die als Primär- und Fremdschlüssel definierten Attribute so zu benennen, dass schon aus der Namensgebung eine Beziehung hervorgeht. Kardinalitäten, d.h. Angaben, wie Entitäten mengenmäßig miteinander zusammenhängen, können in verschiedenen Notationsarten wie der Chen- oder Martin-Notation dargestellt werden (Geisler, 2009, S. 133, 151) (Näheres hierzu siehe Kap. 3.2.2.). Nach diesem ersten Entwurf des ERM muss überprüft werden, ob es folgende Kriterien erfüllt: Vollständigkeit bezüglich der vorgegangenen Anforderungsanalyse, Konsistenzerhaltung, Redundanzfreiheit, Erweiterbarkeit und Verständlichkeit. Sind ein oder mehrere Kriterien nicht erfüllt, muss das ERM mit geeigneten Maßnahmen (z. B. einer tiefergehenden Normalisierung) optimiert werden (Geisler, 2009, S. 301-305).

Parallel zur Konzeption eines Datenmodells wird ein erstes Modell für die Anwendung erstellt, welches sich anfangs auf die Bestimmung der optimalen Methode zur Kommunikation mit der Datenbank und zur Darstellung der Daten beschränkt. Wichtige Fragestellungen sind hierbei, welche grafischen Hilfsmittel dem Anwender für eine möglichst benutzerfreundliche Nutzung zur Verfügung gestellt werden müssen, welcher Art der Darstellung am sinnvollsten ist (Tabellen, Diagramme, etc.) und wie die Schnittstelle zur Kommunikation zwischen Anwender und Datenbank umgesetzt werden kann.

2.2.2. Implementierung von Datenbankanwendungen

Nach der Erstellung eines konzeptionellen Datenmodells in Form eines ERM und der Evaluierung dieses Modells muss für die Implementierung zuerst ein geeignetes Datenbanksystem (DBS) werden (Geisler, 2009, S. 307). Hauptkriterien sollten die technische Eignung und die Kosten des dazu gehörigen Datenbankmanagement-Systems (DBMS) sein, welches als Verwaltungssoftware die Umsetzung des konzeptionellen Modells in die Datenbank überhaupt erst ermöglicht.

Anschließend kann das ERM in ein technisches Datenmodell für das ausgewählte DBS überführt werden (s. Abb. 4). Dabei wird für jeden Entitätstypen eine eigene Datenbanktabelle angelegt, in der eine Zeile einer Entität entspricht (Geisler, 2009, S. 97). In den Tabellen wird für jedes Attribut der Entitäten eine Spalte erstellt und abhängig von der Art der zu speichernden Information und vom Speicherplatzbedarf ein geeigneter Datentyp festgelegt (Geisler, 2009, S. 98). Beziehungen zwischen Entitätstypen werden integriert, indem die Attribute, welche die Entitätstypen miteinander verknüpfen, als Primär- und Fremdschlüssel definiert werden (Geisler, 2009, S. 100 - 105). Somit erhalten die erstellten Tabellen mit ihren Spalten nicht nur Informationen über die gespeicherten Entitäten, sondern auch über die Verknüpfung zu Entitäten anderer Tabellen. Um die Datenbank noch konsistenter zu machen, können für die Attribute neben Primär- und Sekundärschlüssel weitere sog. Constraints (Zwangsbedingungen) definiert werden (Kleinschmidt et al., 2004, S. 86 - 92). Die wichtigsten sind der *Not Null*-Constraint (ein Attribut muss immer einen Wert haben) und der *Unique*-Constraint (ein Attribut muss einzigartig sein). Um die Suchabfragen in der Datenbank zu beschleunigen, können Indizes verwendet werden. Diese bilden eine von der Datenstruktur getrennt gespeicherte und verwaltete Struktur, die schnelle Zugriffe auf einzelne Attributfelder ermöglichen (Geisler, 2009, S. 121 - 122).

Nachdem auch das technische Datenmodell erstellt wurde, kann es nun in dem ausgewählten DBS implementiert werden. Dazu können für den Datenbankentwurf entwickelte Programme verwendet werden, welche oft zusammen mit dem verwendeten DBMS zur Verfügung gestellt werden. Danach kann die Datenbank mit Daten gefüllt werden, welche unter Umständen vorher noch in ein einheitliches und vorteilhaftes Format zu konvertieren und eventuell zu bereinigen sind (Geisler, 2009, S. 309). Schließlich sind das gesamte Datenschema und die gespeicherten Daten mittels einer Datenbeschreibungssprache (DDL: Data Definition Lan-

guage) zu definieren. Dazu wird i. d. R. die international standardisierte Sprache SQL (Structured Query Language) verwendet (Geisler, 2009, S. 65).

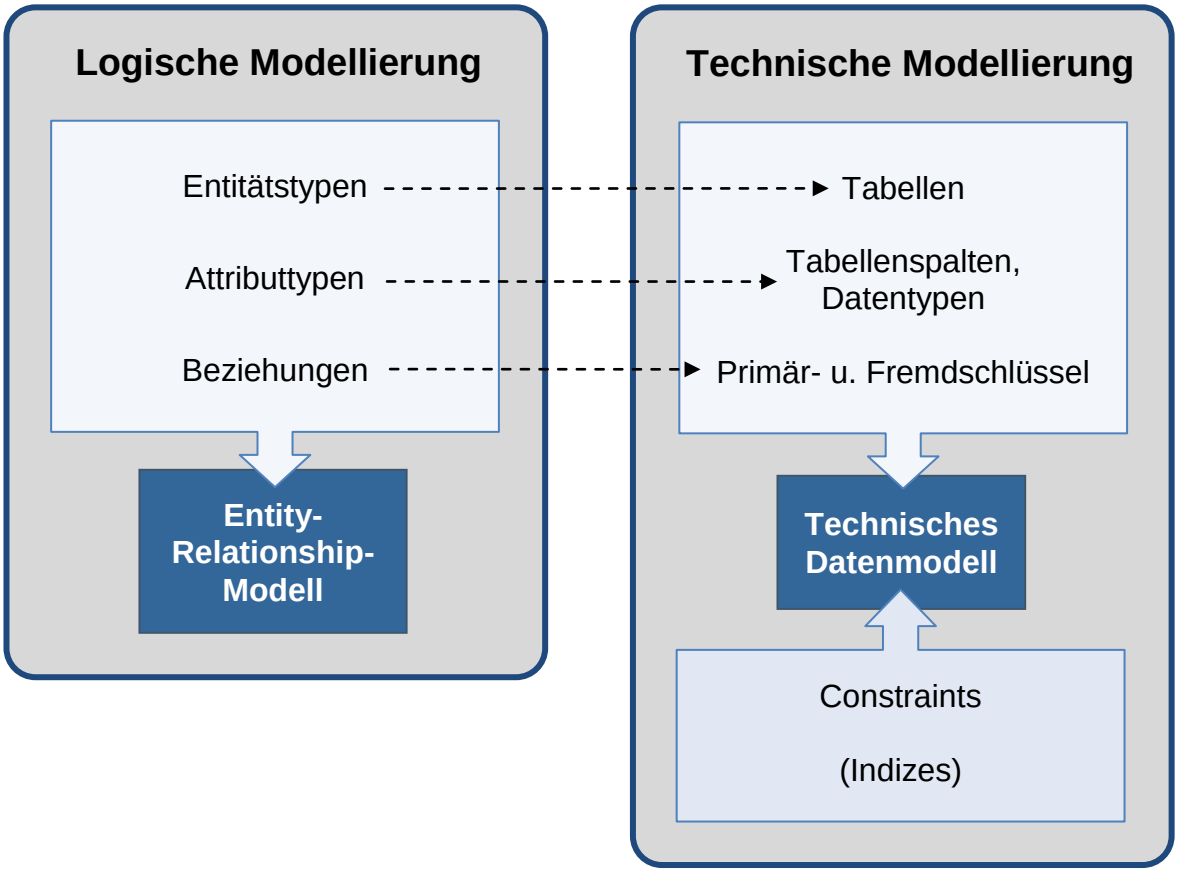


Abb. 4 Verknüpfung der logischen und technischen Datenmodellierung beim Entwurf einer relationalen Datenbank

Beide Tätigkeiten sind wichtige Kernpunkte eines Datenbankentwurfs (s. Abb. 3). Für die logische Modellierung einer relationalen Datenbank wird das Entity-Relationship-Modell (ERM) verwendet, das anschließend auf ein technisches Datenmodell übertragen werden kann. Entitätstypen entsprechen Datenbanktabellen, Attributtypen den Tabellenspalten und Entitätsbeziehungen den Primär- u. Fremdschlüsseln. Bei der technischen Modellierung werden zusätzlich Constraints (Einschränkungen) sowie eventuelle Indizes festgelegt. Das resultierende technische Datenmodell ist Grundlage für die Erstellung und Definition eines Datenschemas in dem ausgewählten Datenbanksystem.

Um die Datenbank auch aus Anwendersicht testen zu können, muss das Anwendungsmodell umgesetzt werden. Dazu werden als erstes geeignete Programmiersprachen ausgewählt, mit deren Hilfe die Anwendung realisiert werden kann. Für die Erstellung einer webbasierten grafischen Benutzeroberfläche haben sich die Webtechnologien HTML (Hypertext Markup Language), XML (Extensible Markup Language) und CSS (Cascading Style Sheets) bewährt (Balzert, 2007, S. 2,44,148). HTML und XML sind Auszeichnungssprachen zur strukturierten

Erstellung von Webseite, die von allen Webbrowsern (Programme zur Darstellung von Webseiten) unterstützt werden, und bieten sich daher für den Aufbau eines Grundgerüstes für die Anwendung an (Balzert, 2007, S. 2,148). Für Benutzereingaben bieten diese Sprachen verschiedene Bedienelemente wie Eingabefelder, Auswahllisten, Auswahlkästen und Knöpfe (Balzert, 2007, S. 33-39). Während HTML und XML in erster Linie für die inhaltliche Gliederung der Webseite verwendet werden, dient die Auszeichnungssprache CSS zur konkreten Darstellung (Farbe, Layout, Schriftart usw.) einzelner Webseitenelemente (Balzert, 2007, S. 44). Um die Nutzerinteraktionen auf der Webseite clientseitig (anwenderseitig) noch vor der serverseitigen Bearbeitung auszuwerten hat sich im Zusammenspiel mit HTML und XML die Skriptsprache JavaScript bewährt (Balzert, 2007, S. 98). Damit werden z. B. Benutzereingaben validiert und Inhalte dynamisch generiert oder verändert (Balzert, 2007, S. 104-125). Für die serverseitige Datenverarbeitung dienen Skriptsprachen wie PHP (Hypertext Preprocessor), Perl oder Java. Mit diesen werden die Anwendungslogik und die Schnittstelle zur Datenbank programmiert, welche die Kommunikation zwischen der Anwendung und der Datenbank vermittelt. Konkret können damit z. B. Benutzereingaben zur Datenanfrage in der Datenbank umgesetzt werden. Als Ergebnis liefert die Schnittstelle die entsprechenden Daten aus der Datenbank, die dann auf der Webseite dem Anwender als Antwort auf seine Eingaben präsentiert werden können.

2.2.3. Evaluierung von Datenbankanwendungen

Nach der Implementierung der Datenbank und des Anwendungsprogrammes müssen diese auf ihre Funktionalität und Performance getestet werden (Geisler, 2009, S. 311). Die Tests der Datenbank können schon während der Entwicklung der Anwendung stattfinden. Dazu genügt anfangs auch ein Anwendungsprototyp, in dem nur die wichtigsten Funktionalitäten wie die Schnittstelle und die Datenpräsentation realisiert sind, um zunächst nur das Zusammenspiel der Anwendung mit der Datenbank zu überprüfen (Geisler, 2009, S. 311). Hat sich die Datenbank als konsistent und die Schnittstelle als valide erwiesen, muss die Anwendung noch hinsichtlich ihrer nutzerseitigen und generellen Anforderungen wie einer übersichtlichen Gestaltung, ansprechenden Präsentation sowie einfachen und verständlichen Bedienung validiert werden.

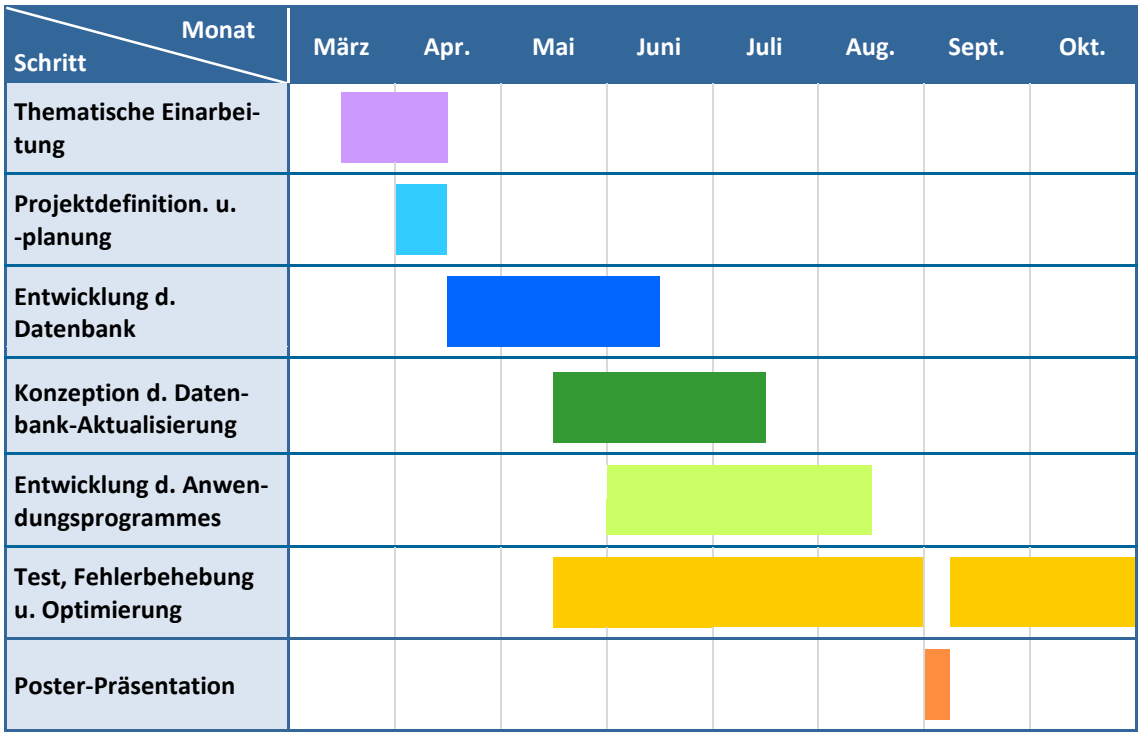
2.2.4. Wartung und Optimierung von Datenbankanwendungen

Die Wartung und Optimierung der Datenbank und des Anwendungsprogrammes finden während der gesamten Laufzeit statt (Geisler, 2009, S. 312). Zum einen muss regelmäßig überprüft werden, ob das Datenbanksystem weiterhin perfomant läuft, insbesondere dann, wenn mit der Zeit immer mehr Daten dazukommen (Geisler, 2009, S. 312). Zum anderen muss für den Fall des Systemausfalls in regelmäßigen Abständen eine Sicherung des Datenmodells und der Daten selbst gemacht werden, damit das System samt Daten auf einen anderen Server migriert werden kann (Geisler, 2009, S. 312). Desweiteren müssen eventuell weitere Entitäten und Attribute angelegt werden, wenn die Anwendung erweitert wird. Für die Optimierung der Webseite ist es sinnvoll, das Design, die Benutzerfreundlichkeit und die Funktionalitäten der Webseite von interessierten Nutzern bewerten zu lassen.

3. MATERIAL & METHODEN

3.1. Arbeitsplan zur Erstellung der Datenbankanwendung PHISTO

Da die Entwicklung der Datenbankanwendung PHISTO verschiedenste strukturelle Schritte umfasste, wurde ein Arbeitsplan erstellt (s. Tab. 2). Die Einteilung der gegebenen Zeit auf die verschiedenen Tätigkeiten verschaffte einen Überblick über die Arbeitsaufwände der einzelnen Schritte, setzte Prioritäten und sicherte die rechtzeitige Realisierung von PHISTO.



Tab. 2 Zeitplan für die Arbeitsschritte bei der Entwicklung der Datenbankanwendung PHISTO

Arbeiten rund um die Datenbank und die Anwendung wurden oft zeitlich parallel getätigt. Ein Vorteil dieser Vorgehensweise lag in der Möglichkeit, die Datenbank durch die gleichzeitige Entwicklung von Teilanwendungen zu validieren und gegebenenfalls frühzeitig Änderungen oder Verbesserungen in der Datenbank vorzunehmen.

Die Einarbeitungsphase diente als Vorbereitung auf die eigentlichen Arbeiten (Schritte 3 - 6) zur Realisierung von PHISTO. Dazu mussten im Rahmen der Aufgabenstellung, eine wissenschaftliche und online verfügbare Datenbankanwendung für PPI pathogenen und humanen Ursprungs zu erstellen, die Grundlagen über PPI-Daten und deren Verwendung in der Bioinformatik erlernt werden. Durch die Analyse anderer PPI-Datenbankanwendungen konnte der wissenschaftliche Stand der Datenverwaltung und -verarbeitung von PPI-Daten bestimmt

werden und die wichtigsten Ziele für die Datenbankanwendung PHISTO definiert werden, nämlich im Vergleich zu diesen umfangreicher, aktueller und benutzerfreundlicher zu sein. Als nächstes wurden bei der Projektdefinition und -planung die Aufgaben und Methoden konkretisiert. Danach folgten die Hauptarbeitsschritte, also die Entwicklung der Datenbank, die Entwicklung eines Konzeptes für die Aktualisierung der Datenbank sowie die Entwicklung der Anwendung. Diese Arbeiten wurden nicht strikt nacheinander ausgeführt, sondern eher zeitgleich, damit die Anforderungen an die Anwendung von vornherein bei der Entwicklung der Datenbank berücksichtigt werden konnten. Zudem konnte die Datenbank durch die gleichzeitige Entwicklung von Teilanwendungen, welche diese verwenden, getestet und validiert werden. Tests, Fehlerbehebungen und Optimierungen wurden parallel zur Implementierung der Datenbank und der Anwendung intensiv durchgeführt und laufen bis heute noch immer. Die Poster-Präsentation diente bei einem Systembiologie-Kongress der Bekanntmachung von PHISTO, um schnell Nutzer für Tests und Feedbacks zu gewinnen.

3.2. Entwicklung der Datenbank für PHISTO

Die Entwicklung der Datenbank erforderte wie in Abb. 5 veranschaulicht mehrere Teilschritte und Daten aus unterschiedlichen Quellen. Zunächst wurden PPI-Daten von neun verschiedenen, externen Datenbanken exportiert und nach Pathogen-Mensch-spezifischen Daten gefiltert. Danach wurden diese nach verschiedenen Kriterien analysiert, um zu entscheiden, ob die Daten zu bearbeiten und zu erweitern sind. Im nachfolgenden Schritt wurden die Daten vereinheitlicht und von Redundanzen befreit (s. Kap. 3.1.1.). Für die Erweiterung der PPI-Daten wurden drei weitere, externe Datenbanken herangezogen, die weltweit als De-Facto-Standard für die Bezeichnung und Beschreibung der jeweils gespeicherten Daten gelten (s. Kap. 3.1.1.). Die ausführliche Analyse, Anpassung und Erweiterung der gesammelten PPI-Daten lieferte eine Zusammenstellung, einheitlicher und nicht redundanter Datensätze aus Kerndaten, welche die PPI eindeutig bestimmen, und aus Metadaten, welche die PPI ergänzend beschreiben. Diese Daten dienten als Grundlage für die Erstellung eines relationalen Datenmodells für PHISTO (s. Kap. 3.1.2.) und zum Füllen der Datenbank nach dessen Implementierung(s. Kap. 3.1.3.).

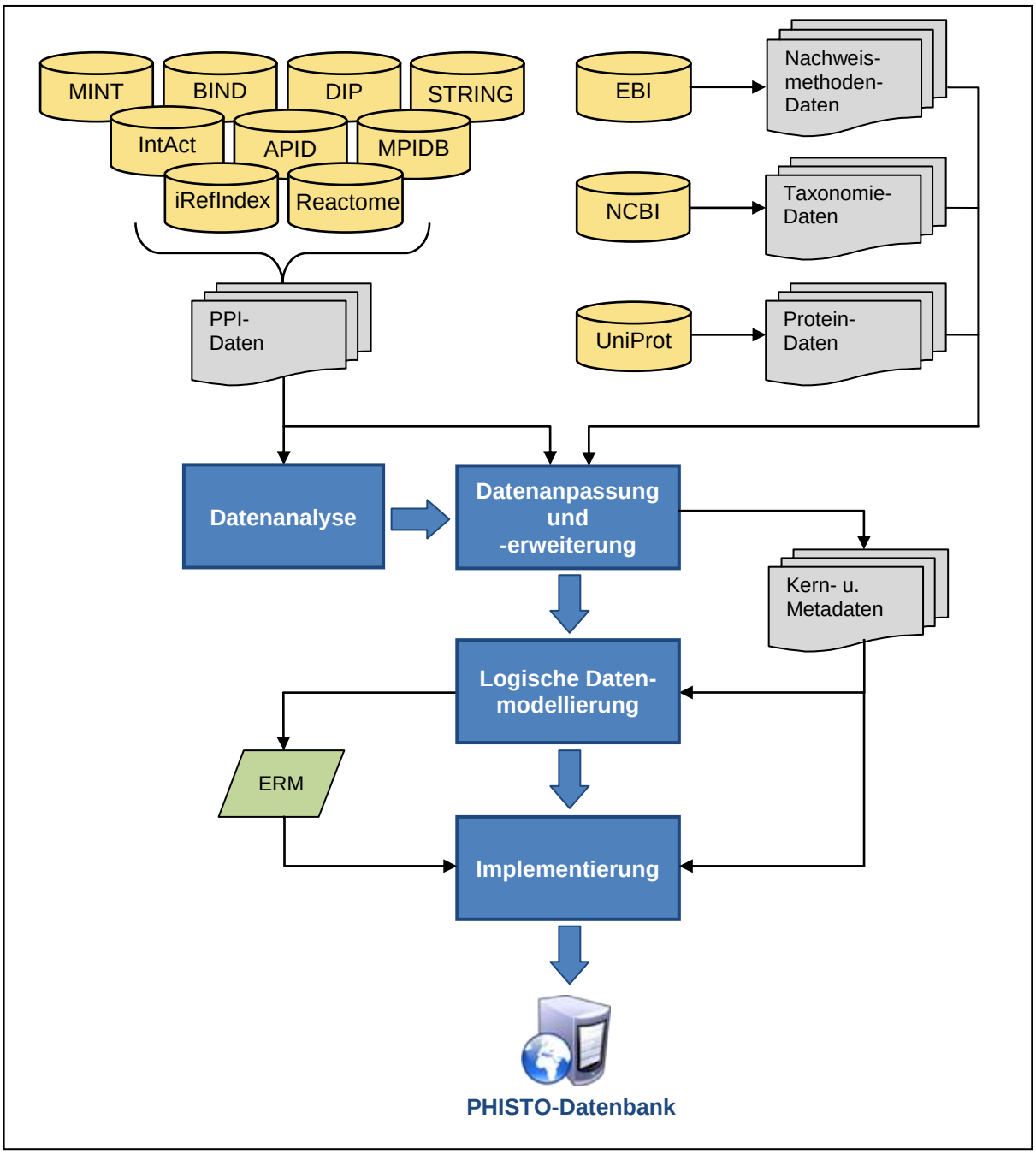


Abb. 5 **Arbeits- und Datenflussdiagramm für die Entwicklung der PHISTO-Datenbank**

Die blau hinterlegten Rechtecke sind Arbeitsschritte der Datenbankenwicklung. Die gelben Zylinder symbolisieren die verwendeten externen Datenbanken. Graue Symbole stellen Daten dar, die entweder von den externen Datenbanken zur Verarbeitung extrahiert wurden oder im Falle der Kern- und Metadaten als Ergebnis der Datenverarbeitung entstanden. Zunächst wurden von neun verschiedenen Datenbanken Mensch-Pathogen spezifische PPI-Daten exportiert, zusammengeführt und analysiert. Im nächsten Schritt wurden die Daten vereinheitlicht und um Metadaten ergänzt. Dazu wurden Protein-, Taxonomie- und Nachweismethodendaten aus drei weiteren Datenbanken exportiert, die weltweit als De-Facto-Standard für die Bezeichnung und Beschreibung der jeweils gespeicherten Daten gelten. Mit den so erhaltenen Kern- und Metadaten der PPI konnte ein logisches Datenmodell erstellt werden (grün), das als Grundlage für das technische Datenmodell bei der Implementierung diente. Schließlich wurden die PPI-Daten mittels des relationalen Datenbankverwaltungssystems MySQL in eine Datenbank importiert.

3.2.1. Analyse, Anpassung und Erweiterung der gesammelten PPI-Daten

Die in der PHISTO-Datenbank gespeicherten Daten sind den folgenden neun online verfügbaren PPI-Datenbanken entnommen: APID, BIND, DIP, IntAct, iRefIndex, MINT, MPIDB, Reactome and STRING. Es sei darauf hingewiesen, dass in Kapitel 2.1. drei weitere Datenbanken aufgeführt sind, die für PHISTO nicht verwendet wurden: BioGrid, PATRIC und HPIDB. BioGrid und PATRIC wurden erst nach den praktischen Tätigkeiten für PHISTO entdeckt, allerdings sind die PPI-Daten von BioGrid in den Daten der sekundären Datenbanken iRefindex und APID enthalten (s. Tab. 1). HPIDB konnte nicht berücksichtigt werden, weil die Daten dieser Datenbank nicht runtergeladen werden konnten.

Die Daten jeder der verwendeten Datenbanken wurden in eine eigene Datei mit einem Format für tabellarisch strukturierte Daten exportiert. In solch einer Datei entspricht eine Zeile einem einzelnen Datensatz aus mehreren Datenelementen, die wiederum spaltenweise angegeben werden. In diesem Fall entsprach eine Zeile immer einer einzelnen PPI-Datensatz, allerdings handelte es sich bei dem Großteil der Datensätze nicht um PPI zwischen Pathogenen und dem menschlichen Organismus, sodass diese als erstes herausgefiltert werden mussten. Dazu wurde in Microsoft Excel ein Algorithmus geschrieben, der alle Datensätze löscht, welche die folgenden Kriterien nicht erfüllen:

- Das eine Protein muss pathogenen Ursprungs sein.
- Das andere Protein muss humanen Ursprungs sein.

Desweiteren unterschieden sich die Informationen, die pro PPI-Datensatz angegeben sind, von Datenbank zu Datenbank und teils auch innerhalb ein und derselben Datenbank. Zum einen unterschied sich die Art der Informationen für eine PPI wie in Tab. 3 beispielhaft dargestellt und zum anderen war ein und dieselbe Informationsart wie in Tab. 4 beispielhaft dargestellt unterschiedlich angegeben. Daher mussten die Datensätze vor ihrer Speicherung in die PHISTO-Datenbank zunächst in ein einheitliches und konsistentes Format gebracht werden. Dazu wurde der von der Proteomics Standard Initiative (PSI), einer Arbeitsgruppe der Human Proteome Organisation (HUPO) definierte Standard für die mindestens benötigten Informationen zum Veröffentlichen von molekularen Interaktionen herangezogen, der inzwischen von den meisten Wissenschaftlern und vielen primären Datenbanken der molekularen Interaktomforschung wie BIND, DIP, IntAct, MINT und MPact verwendet wird (Orchard et al., 2007).

Datenbank	IntAct	MPIDB
ID-Nr. des Proteins A	x	x
ID-Nr. des Proteins B	x	x
Alternative ID-Nrn. des Proteins A	x	
Alternative ID-Nrn. des Proteins B	x	
Bezeichnung des Proteins A	x	x
Bezeichnung des Proteins B	x	x
ID-Nr. des Gens, das Protein A kodiert		x
ID-Nr. des Gens, das Protein B kodiert		x
Nachweismethode	x	x
Autor der Publikation		x
ID-Nr. der Publikation	x	x
ID-Nr. des Organismus A	x	x
ID-Nr. des Organismus B	x	x
ID-Nr. der Art der Interaktion	x	x
Anzahl der Belege für die PPI		x

Tab. 3 **Beispiel für die Unterschiede bei der Anzahl der Informationsangaben zwischen verschiedenen PPI-Datenbanken**

Beispielhaft werden die Unterschiede zwischen den Datenbanken IntAct und MPIDB für die unterschiedlichen Informationsangaben zu einer PPI anhand der wichtigsten Attribute gezeigt. Während bei IntAct alternative ID-Nrn. für die interagierenden Proteine gespeichert sind, fehlt diese Information bei MPIDB gänzlich. Dafür führt MPIDB Informationen über die ID-Nrn. der Gene, welche die Proteine kodieren, über den Autor der Publikation, in der die PPI veröffentlicht wurde, und über die Anzahl der Belege für die PPI.

Datenbank	Eintrag für die Bezeichnung des Proteins	Eintrag für die taxonomische ID-Nr. des Organismus
IntAct	psi-mi:2b14_human(display_long)|uniprotkb:HLA-DRB1(gene name)|psi-mi:HLA-DRB1(display_short)|uniprotkb:MHC class II antigen DRB1*4(gene name synonym)	taxid:9606(human)|taxid:9606(Homo sapiens)
MPIDB	uniprotkb:HLA-DRB1	taxid:9606

Tab. 4 Beispiel für die Unterschiede bei den Informationsangaben zwischen verschiedenen PPI-Datenbanken

Beispielhaft werden die Unterschiede zwischen den Datenbanken IntAct und MPIDB bei zwei Informationsangaben gezeigt. Während bei IntAct viele verschiedene Bezeichnungen für ein interagierendes Protein existieren, verwendet MPIDB eine einzige Kurzbezeichnung. Die taxonomische ID-Nr. wird von beiden Datenbanken wie in diesem Beispiel mit 9606 angegeben, doch IntAct speichert darüber hinaus auch die englische Bezeichnung (human) und den lateinischen Namen aus der biologischen Systematik (Homo sapiens).

Nach diesem Standard sollten PPI-Datensätze idealerweise mindestens folgende Informationen beinhalten:

- Eindeutige, international geltende ID-Nr. für die interagierenden Proteine
- Eine eindeutige, international geltende ID-Nr. für die experimentelle Methode, mit der die Interaktion nachgewiesen wurde
- Eine eindeutige, international geltende ID-Nr. für den Organismus, in dem die Interaktion nachgewiesen wurde
- Die Publikation, in der die Interaktion veröffentlich wurde

Darüber hinaus wird die Angabe weiterer Informationen empfohlen, unter anderem einer ID-Nr. für die Proteinnachweis-Methode, einer alternative ID-Nr. für die interagierenden Proteine sowie von Bezeichnungen der Proteine.

Für die PHISTO-Datenbank wurde festgelegt, dass jeder PPI-Datensatz folgende Kerninformationen beinhalten muss, damit er einerseits den HUPO-PSI-Standard erfüllt und andererseits eindeutig zu identifizieren ist:

- Die UniProt-ID-Nr. für die interagierenden Proteine: eine von UniProt (Universal Proteine Resource) festgelegte Kennung für Proteine
- Die Tax-ID-Nr. für die pathogenen Organismen: eine von NCBI (National Center for Biotechnology) festgelegte Kennung für taxonomische Zuordnung von Organismen

- Die MI-ID-Nr. für die experimentelle Methode, mit der die Interaktion nachgewiesen wurde: eine vom PSI festgelegte Kennung für Methode, mit denen molekulare Interaktionen nachgewiesen werden
- Die PubMed-ID-Nr. für die Publikation, in der die Interaktion veröffentlich wurde: eine von NCBI festgelegte ID-Nr. für wissenschaftliche Publikationen

Diese Kerninformationen sind allerdings sehr abstrakt und bieten dem Nutzer keine Möglichkeit, die Datenbank nach Begriffen wie z. B. „HIV“ oder „Tyrosinkinase“ zu durchsuchen. Daher wurde festgelegt, dass die Kerninformationen um folgende zusätzliche, beschreibende Informationen zu erweitern sind:

- Bezeichnungen für die interagierenden Proteine
- Taxonomische Informationen zu dem pathogenen Organismus, dessen Protein eine Interaktion mit humanen Proteinen aufweist
- Bezeichnung der experimentellen Methode zum Nachweis der Interaktion

Die für die PHISTO-Datenbank festgelegten Kerninformationen waren in dem Großteil aller gesammelten PPI-Datensätze enthalten, sodass für diese Datensätze lediglich die redundanten Kerninformationen entfernt werden mussten, um die Datensätze aller Datenbanken in ein einheitliches Format zu bringen. Datensätze, in denen eine Kerninformation fehlte, wurden in dieser Arbeit ausgeschlossen. Dabei handelte es sich in allen Fällen um eine fehlende Angabe der MI-ID-Nr. für die experimentelle Methode. Inzwischen wurde von anderen Entwicklern ein Algorithmus geschrieben, der für solche Datensätze automatische die fehlende MI-ID-Nr. ergänzt, indem mit Hilfe anderer Datenfelder im Internet nach der zugehörigen Publikation und innerhalb dieser nach der MI-ID-Nr. gesucht wird, wodurch inzwischen auch diese vorerst unberücksichtigten PPI in die Datenbank aufgenommen wurden (Durmus Tekir S. , 2013, S. 27). Nach Vereinheitlichung der PPI-Datensätze aus den neun verschiedenen Datenbanken wurde diese in einer einzelnen Datei zusammengeführt. Da in den Datenbanken teilweise dieselben PPI gespeichert waren, mussten nach der Zusammenführung der Daten noch die doppelten Einträge gelöscht werden. Die zusätzlichen, beschreibenden Informationen hätten nicht oder nur teilweise aus den gesammelten PPI-Datensätzen bezogen werden können. Daher wurden stattdessen wie von HUPO-PSI empfohlen für diese Daten die drei öffentlich zugänglichen Datenbanken von UniProt (http://www.uniprot.org/), NCBI (http://www.ncbi.nlm.nih.gov/) und EBI (http://www.ebi.ac.uk/) hinzugezogen. Die UniProt-Datenbank sammelt umfassende

Informationen zu Proteinen. Von dieser wurden international gängige, englischsprachige Bezeichnungen für die interagierenden Proteine bezogen. Die NCBI-Datenbank stellt chemische, biochemische, molekularbiologische, aber auch taxonomische Daten zur Verfügung. Von dieser wurden taxonomische Informationen bezogen. Die EBI-Datenbank bietet biochemische und molekularbiologische Daten an. Von dieser wurden die Bezeichnungen der experimentellen Methoden zum Nachweis der Interaktion bezogen. Die Vorgehensweise des Datenexports aus diesen Datenbanken wird nachfolgend anhand der Taxonomie-Informationen erläutert: Zuerst wurden alle Tax-ID-Nrn. der pathogenen Organismen aus den PPI-Datensätzen extrahiert. Anschließend wurde mit diesen Daten bei der NCBI-Datenbank angefragt, wie die zugehörigen Bezeichnungen des Reiches, der Familie und des Stammes der Pathogene lauten. Diese Informationen wurden runtergeladen und zusammen mit den zugehörigen Tax-ID-Nrn. in getrennten Tabellen gespeichert. Insgesamt erhielt man so drei voneinander getrennte Tabellen für die Proteinbezeichnungen, die taxonomischen Informationen und für die Bezeichnungen der experimentellen Methoden zur Aufklärung der PPI, die wie in Abb. 6 veranschaulicht über die ID-Nrn. mit den PPI-Datensätzen in der vierten Tabelle verknüpft sind.

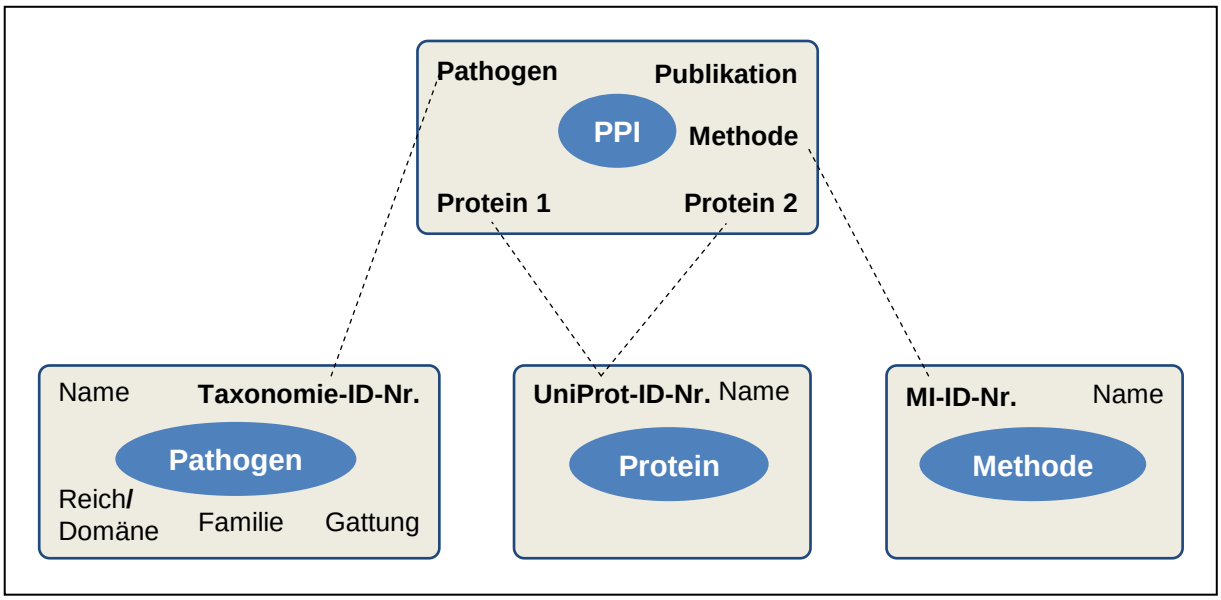


Abb. 6 Verknüpfung von Kern- und Metadaten der PPI

Oben ist die PPI mit den Kerndaten zu sehen. Unten sind die drei aus Datensicht eigenständigen Objekte Pathogen, Protein und Nachweismethode mit ihren Metadaten zu sehen, welche die PPI näher beschreiben. Über ID-Nrn. sind die Metadaten mit den Kerndaten der PPI verknüpft. Diese ID-Nrn. können wie hier schon ansatzweise dargestellt für die Erstellung eines relationalen Datenmodells verwendet werden, in dem die verschiedenen Entitätstypen (blaue Ellipsen) über gemeinsame Attribute verknüpft sind. Die Metadaten beschränken sich zunächst auf die für die Suche und Ergebnisdarstellung wichtigsten Daten, da auf weitere Daten über Links zu den referierten Datenbanken zugegriffen werden kann (s. Abb. 14).

3.2.2. Datenmodellierung für die PHISTO-Datenbank

Die logische Datenmodellierung in Form eines ERM dient der Vorbereitung zur technischen Umsetzung in einem DBMS. Grundlage für die Erstellung des ERM waren die in Kap. 3.1.1. erstellten Tabellen, welche jeweils einen Entitätstypen darstellen, und die Verknüpfungen dieser Tabellen (s. Abb. 6). Zunächst wurde für diejenigen Objekte, welche die PPI mit ihren Metadaten beschreiben, also für die Proteine, die Pathogene und die Nachweismethoden, ein eigener Entitätstyp angelegt (s. Abb. 7). Dabei erhielten die Entitätstypen diejenigen Attribute, die auch in den vorbereiteten Tabellen enthalten sind, beispielsweise im Falle der Pathogene die Tax-ID-Nr. als *idTaxonomy*, die taxonomische Art/Spezies oder falls gegeben den Stamm als *Name*, die Familie als *Family* und das Reich bzw. die Domäne als *Kingdom*. Allen Attributen wurden anschließend sinnvolle Datentypen zugewiesen, z. B. bei idTaxonomy der Typ *int(10)*, was bedeutet, dass dieses Attribut vom Datentyp Integer (ganzzahliger Wert) ist und maximal 10 Zeichen lang sein darf. Dann wurde in jeder der drei Entitätstypen dasjenige Attribut als Primärschlüssel definiert, welches das jeweilige Objekt eindeutig identifiziert und eine Verknüpfung mit PPI-Objekten ermöglicht, also beim Beispiel der Pathogene die idTaxonomy. Gekennzeichnet ist ein Primärschlüssel durch die angehängte Abkürzung *PK* (Primary Key).

Nach der Modellierung der drei Entitätstypen für die Metadaten wurde ein eigener Entitätstyp für die PPI erstellt. Als Attribute besitzen die PPI bis auf die PubMed-ID-Nr., hier *PubMedId*, nur Fremdschlüssel, die auf die jeweiligen Primärschlüssel der anderen Entitäten verweisen. Gekennzeichnet ist ein Sekundärschlüssel durch die angehängte Abkürzung *FK* (Foreign Key). Für eine schnelle Erkennbarkeit der Beziehungen zwischen Fremd- und Primärschlüsseln in der später erzeugten Datenbank wurden alle Fremdschlüssel mit einem vorangehenden *fk* und anknüpfend mit der Bezeichnung des verwiesenen Primärschlüssels benannt. Z. B. verweist der Fremdschlüssel *fkIdTaxonomy* im Entitätstyp *PPIs* auf den Primärschlüssel *idTaxonomy* im Entitätstyp *Pathogens*. Insgesamt besitzt der Entitätstyp PPIs fünf Attribute: *fkIdUniprotHP* für das interagierende humane Protein, *fkIdUniprotPP* für das interagierende Protein pathogenen Ursprungs, *fkIdTaxonomy* für den pathogenen Organismus, *fkIdMI* für die experimentelle Nachweismethode und *PubMedId* für die Publikation, in der die PPI veröffentlicht wurde. Da es kein einzelnes Attribut besitzt, welches als Primärschlüssel definiert werden kann, wurde die Gesamtheit aller Attribute als Primärschlüssel definiert. Das bedeutet, dass ein PPI-Datensatz erst durch die Kombination aller seiner Attributwerte eindeutig wird.

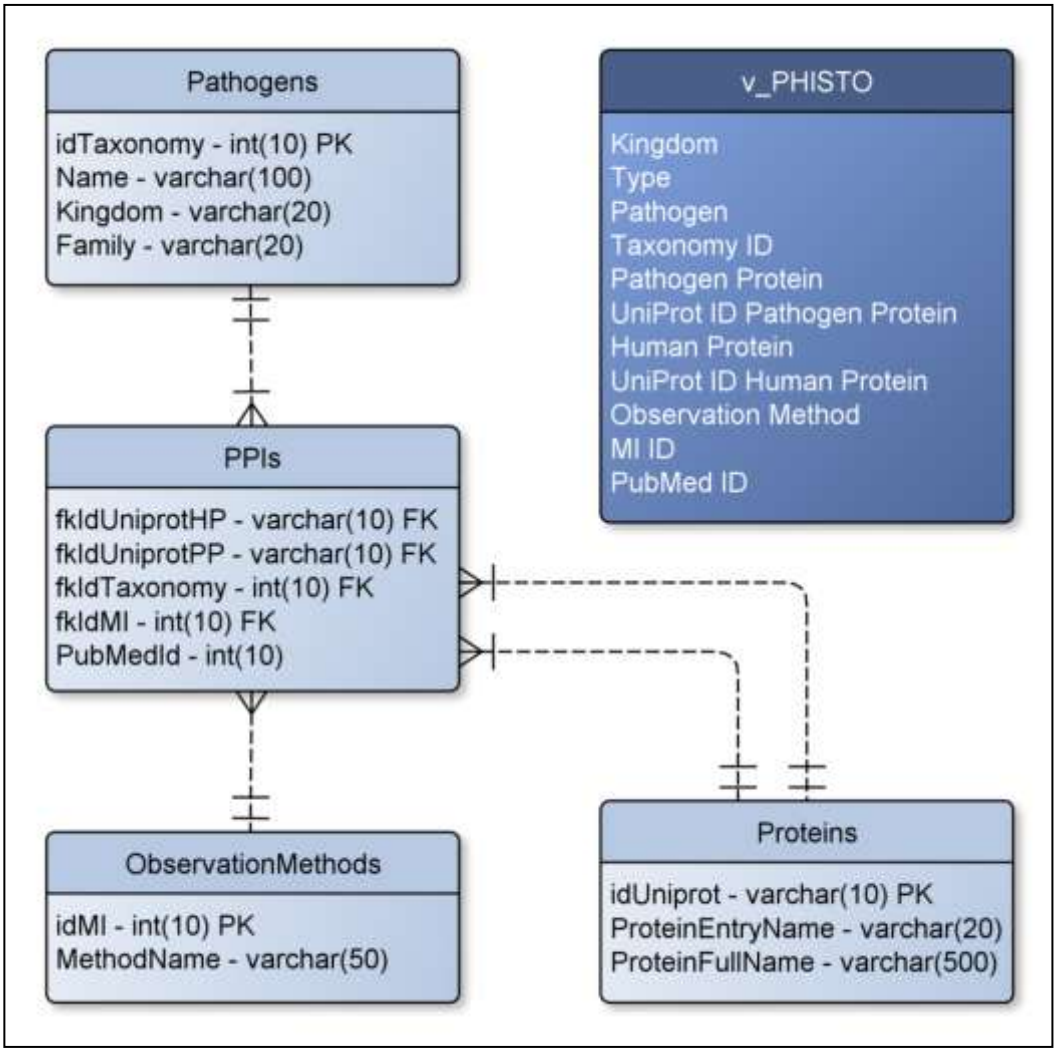


Abb. 7 Entity-Relationship-Modell der PHISTO-Datenbank

Alle Entitätstypen sind durch hellblaue Rechtecke dargestellt. Die Kerndaten der PPIs befinden sich in dem Entitätstyp *PPIs*, dessen Attribute bis auf die PubMed ID-Nr. Fremdschlüssel (fkID…) sind. Wie in der hier dargestellten Martin-Notation (gestrichelte Linien) sind diese über die Primärschlüssel (id…) über *1:n*-Beziehungen mit den anderen Entitätstypen *Pathogens*, *Proteins* und *ObservationMethods* verknüpft, welche die Metadaten enthalten. Als Schlüssel werden die UniProt-ID-Nr. für die Proteine, die Taxonomie-ID-Nr. für die Pathogene und die MI-ID-Nr. für die Aufklärungsmethode verwendet (s. auch Abb. 6). Der Entitätstyp *PPIs* besitzt als Primärschlüssel die Kombination aus allen Attributen. Aus diesen vier Entitätstypen soll die Sicht *v_PHISTO* erstellt werden, welche für die Ergebnis-Darstellung auf der Webseite verwendet wird (s. Kap 3.2.3.).

Begriffe & Abkürzungen

int: ganzzahlige Werte; varchar: beliebige Zeichenkette; PK: Primary Key; FK: Foreign Key

Die Beziehungen, also die Verknüpfungen zwischen den Entitäten über ihre Primär- und Fremdschlüssel, sind in der Martin-Notation dargestellt. Jede gestrichelte Linie zwischen zwei Entitätstypen bedeutet eine Beziehung. Die Kardinalitäten, also die mengenmäßigen Bezie-

hungen der Entitäten, sind in der Martin-Notation durch Symbole an den Verbindungsstellen angegeben. Hier sind alle Beziehungen vom Typ *1 : n*. Das bedeutet, dass jede Entität des einen Entitätstyps mit beliebig vielen Entitäten des anderen Entitätstyps in Beziehung stehen kann. Für die Proteine veranschaulicht bedeutet dies, dass jedes Protein in der Tabelle Proteins nur einmal und in der Tabelle PPIs mehrfach vorkommen kann.

3.2.3. Implementierung des Datenmodells in die PHISTO-Datenbank

Als DBMS, in dem das Datenmodell implementiert wurde, fiel die Wahl auf das weit verbreitete und frei verfügbare System MySQL. Als grafisches Administrationsprogramm für MySQL wurde phpMyAdmin verwendet. Als erstes wurde in MySQL für das gesamte Datenmodell ein neues Datenbankschema erstellt. Dieses dient als getrennt verwalteter Bereich für die Realisierung des Datenmodells. Anschließend wurde dem ERM gemäß die technische Datenmodellierung vorgenommen (s. Abb. 7).

Zunächst wurde für die drei Entitätstypen Pathogens, Proteins und ObservationMethods jeweils eine eigene Tabelle mit den jeweiligen Attributen als Tabellenspalten angelegt. Dabei erhielten die Spalten die im ERM definierten Datentypen. Spalten, die Kennungen speichern, wurden wie im ERM festgelegt als Primärschlüssel definiert. Dann wurde für den Entitätstyp PPIs eine vierte Tabelle angelegt. Hier wurde jede Tabellenspalte, die auf einen Primärschlüssel verweist, als Fremdschlüssel definiert. Dies führt in dem DBMS dazu, dass ein Fremdschlüsselwert erst dann in der Datenbank existieren kann, wenn dieser Wert auch als Primärschlüssel existiert. Das bedeutet, dass für die Speicherung eines PPI-Datensatz immer auch die entsprechenden Metadaten existieren müssen. Dies gewährleistet die vollständige und konsistente Speicherung von PPI.

Nach der Erstellung der vier Datenbanktabellen konnte der Datenimport vorgenommen werden. Zuerst mussten die Tabellen Pathogens, Proteins und ObservationMethods gefüllt werden. Dann erst konnte die Tabelle PPIs gefüllt werden, da deren Fremdschlüsselwerte auf die Primärschlüsselwerte der anderen Tabellen verweisen.

Zusätzlich zu den Tabellen wurde die Sicht *v_PHISTO* erstellt (s. Abb. 7, dunkelblauer Kasten). Eine Sicht ist eine Zusammenstellung mehrerer Tabellen, welche durch eine Abfrage in

dem DBMS definiert wird und wie eine normale Datenbanktabelle abgefragt werden kann. In diesem Fall verknüpft die Sicht alle beschreibenden Metadaten der drei Tabellen Pathogens, Proteins und ObservationMethods über ihre Beziehungen zu der Tabelle PPIs, sodass in ihr die abstrakten PPI-Informationen um anschaulichere Informationen erweitert werden. Die Sicht v_PHISTO wird für die gesamten nutzerseitigen Anfragen und die Datenpräsentation auf der Webseite verwendet.

3.3. Aktualisierung der PPI-Daten in der PHISTO-Datenbank

Die in der PHISTO-Datenbank gespeicherten PPI-Daten können nach einiger Zeit überholt sein, wenn sich gewisse Informationen ändern. Am wahrscheinlichsten sind Änderungen bei den Proteinen. Zum einen können sich Bezeichnungen ändern und zum anderen können sich die UniProt-ID-Nrn. ändern. Die Änderung der UniProt-ID-Nrn. beruht laut UniProt auf folgenden Mechanismen:

- Eine bestehende Nr. wird durch eine neue, primär gültige Nr. ersetzt (*replaced*)
- Eine bestehende Nr. wird in zwei neue, primär gültige Nrn. geteilt, welche jeweils ein anderes Protein repräsentieren (*demerged*)
- Eine bestehende Nr. wird mit einer anderen Nr. zu einer neuen Nummer zusammengefasst, bleibt aber die primär gültige Nr. (*merged*)

Im ersten und zweiten Fall müssen die Änderungen auch in der PHISTO-Datenbank berücksichtigt werden, damit zu jedem Zeitpunkt nur die primären UniProt-ID-Nrn. gespeichert sind. Dazu wurde das in Abb. 8 veranschaulichte Aktualisierungskonzept entwickelt. Zuerst werden die UniProt-ID-Nrn. aller Proteine aus der PHISTO-Datenbank extrahiert. Mit diesen wird bei der UniProt-Datenbank eine Anfrage nach den Informationen inklusive der Änderungen zu den Proteinen gemacht. Diese Informationen werden in einer Datei gespeichert und dienen als Eingabe für ein mit Perl geschriebenes Programm, welches die Informationen aus UniProt mit den Informationen in der PHISTO-Datenbank vergleicht und bei Unterschieden die Daten in der PHISTO-Datenbank aktualisiert. Zum jetzigen Standpunkt werden von dem Programm keine ID-Nrn. berücksichtigt, welche in zwei neue Nummern geteilt sind, weil nicht automatisch bestimmt werden kann, mit welches der beiden Proteine für die PPI übernommen werden muss. Alternativ muss dies recherchiert und die Aktualisierung in der Datenbank manuell vorgenommen werden.

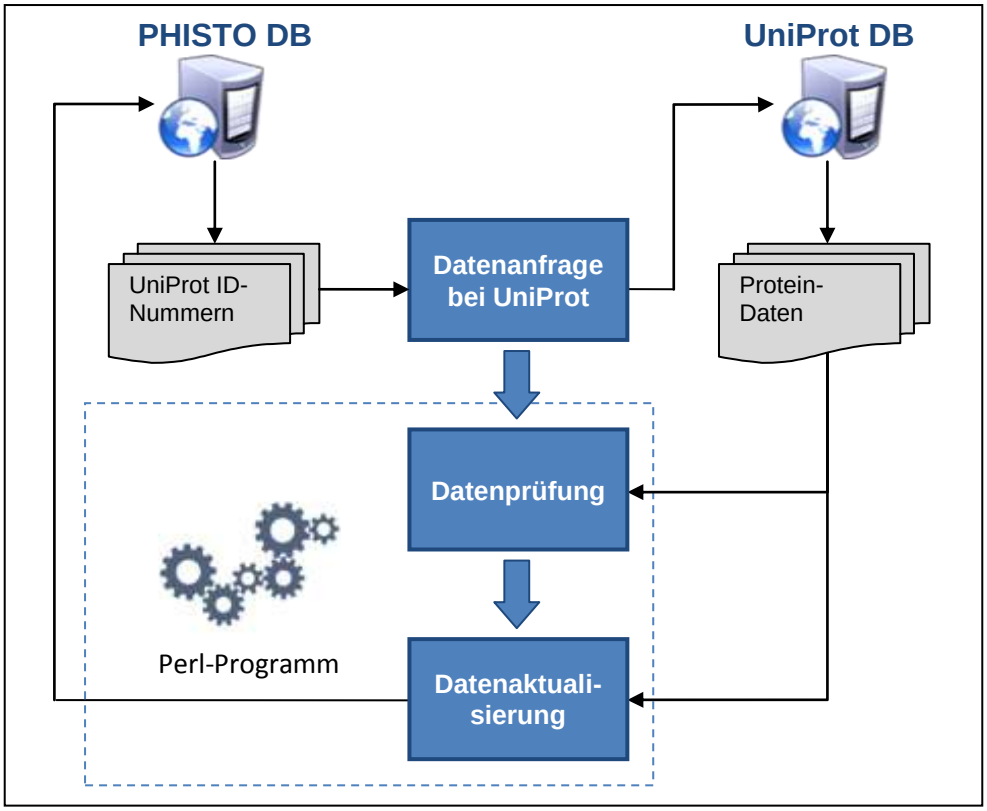


Abb. 8 **Aktualisierungsmethode für die PHISTO-Datenbank**

Zuerst werden alle UniProt-ID-Nummern aus der Tabelle *Proteins* (s. Abb. 7) exportiert, um damit eine Datenanfrage bei der UniProt-Datenbank zu machen. Die erhaltenen Protein-Daten werden exportiert und mit den Daten in der PHISTO-Datenbank verglichen. Bei Unterschieden in den Protein-Daten zwischen der PHISTO-Datenbank und der UniProt-Datenbank werden die Daten in der PHISTO-Datenbank aktualisiert. Die beiden Schritte der Datenprüfung und Datenaktualisierung werden automatisiert von einem Perl-Programm ausgeführt.

3.4. Entwicklung der Benutzeranwendung für PHISTO

3.4.1. Definition der Benutzeranwendung

Bevor mit der Programmierung der Benutzeranwendung für die PHISTO-Datenbank begonnen werden konnte, mussten die Anforderungen erörtert und definiert werden. Die gegebenen Daten und die Thematik dieser Arbeit gaben zwar vor, dass den Nutzern eine grafische Oberfläche angeboten werden muss, welche in erster Linie Recherchen in der Datenbank erlaubt, doch für die Details musste recherchiert werden, was darüber hinaus die wichtigsten Kriterien für eine ausreichende Funktionalität und Benutzerfreundlichkeit sind. Dazu wurde die Webseite der IntAct-Datenbank für PPI als Referenz herangezogen, da diese seit 2004 weltweit öffentlich genutzt wird und inzwischen auch in der Literatur zu PPI fest verankert ist. Die

Analyse dieser Webseite ergab, dass folgenden Anforderungen bei der Entwicklung der Benutzeranwendung die höchste Priorität eingeräumt werden muss:

- Die Benutzeranwendung muss weltweit über das Internet in Form einer Webseite erreichbar sein.
- In der Datenbank müssen PPI über eine einfache und schnelle sowie über eine erweiterte Suche recherchiert werden können. Die erweiterte Suche muss genügend Möglichkeiten bieten, um mehrere und möglichst genaue Suchkriterien zu erstellen.
- Recherchen müssen auch ohne die Angabe eines Suchausdrucks möglich sein, und zwar anhand der taxonomischen Informationen zu den Pathogenen („Durchstöbern“).
- Rechercheergebnisse müssen in übersichtlicher, tabellarischer Form dargestellt werden und mit Verweisen zu externen Datenbanken versehen werden.
- Rechercheergebnisse müssen in verschiedenen tabellarischen Formaten runtergeladen werden können.
- Statistiken über die in der Datenbank gespeicherten Daten müssen schnell einsehbar sein.
- Die Navigation durch die Hauptfunktionalitäten muss über ein übersichtliches Menü realisiert werden, welches bei Bedarf einfach erweiterbar ist.

3.4.2. Implementierung der Benutzeranwendung

Die Anwendung wurde als Webseite mit den Webtechnologien PHP, JavaScript, HTML und CSS entwickelt, welche von den Benutzern interaktiv verwendet werden kann und dementsprechend dynamisch aufgebaut werden muss. Als Entwicklungsumgebung, mit der das Skript für die Webseite geschrieben wurde, diente die frei verfügbare Software Notepad++. Zum Testen der Anwendung wurde lokal ein frei verfügbarer Webserver von Apache installiert, welcher das erstellte Skript serverseitig und dynamisch interpretiert und daraus ein Dokument erstellt, das dann als Seite in einem Webbrowser dargestellt wird. Bei der Erstellung des Skripts wurde auf eine strukturierte und gut lesbare Schreibweise wertgelegt. Dies bedeutet, dass zum einen logisch oder inhaltlich voneinander getrennte Bereiche gekennzeichnet wurden und zum anderen bei der Benennung von Variablen und Funktionen selbstsprechende Bezeichnungen verwendet wurden.

Weil die zu entwickelnde Webseite und die beinhalteten Funktionalitäten überschaubar waren, wurde das Skript prozedural programmiert, also als eine Abfolge von Anweisungen, die von verschiedenen Kontrollstrukturen gesteuert werden und in voneinander abgegrenzte, logische Teile (Funktionen) zerlegt sind. In Abb. 9 ist ein Flussdiagramm zu sehen, dass die prozeduralen und kontrollierten Aufruf der wichtigsten Funktionen verdeutlicht.

Der Kopfbereich der Webseite mit dem Navigationsmenü (s. Abb. 10) und der Fußbereich mit einigen rein informativen Angaben sind statisch, also in der Darstellungsweise nicht veränderbar, sodass für diese Bereiche jeweils nur eine einzige Funktion aufgerufen werden muss, und zwar am Anfang des Skriptes für den Kopfbereich und am Ende für den Fußbereich. Der dazwischen liegende Seitenkörper, also der mittlere Bereich zur Anzeige der Funktionalitäten sowie der Rechercheergebnisse, wird dynamisch durch verschiedene Funktionen aufgebaut, deren Aufruf von einer übergeordneten Kontrollfunktion gesteuert wird (gestrichelte Linie). Dabei sind für die meisten Funktionalitäten wie beispielweise die Anzeige der Statistiken eigene Funktionen verantwortlich, welche der Übersichtlichkeit halber wiederum in mehrere Unterfunktionen aufgeteilt sind. Für Funktionalitäten, die einen ähnlichen Zweck erfüllen und daher auch ähnliche Anweisungen benötigen, wurden dagegen die benötigten Funktionen nur einmal geschrieben, z. B. bei der Datenrecherche mit einem einzelnen Suchausdruck und der Datenrecherche mit mehreren Suchkriterien. Dadurch konnten mehrfach benötigte Anweisungen wie in diesem Beispiel die Verbindung zur Datenbank für Datenabfragen wiederverwendet und somit redundanter Quellcode vermieden werden. Eine Unterscheidung wird funktionsintern über Parameter vorgenommen, die an die Funktion übergeben werden.

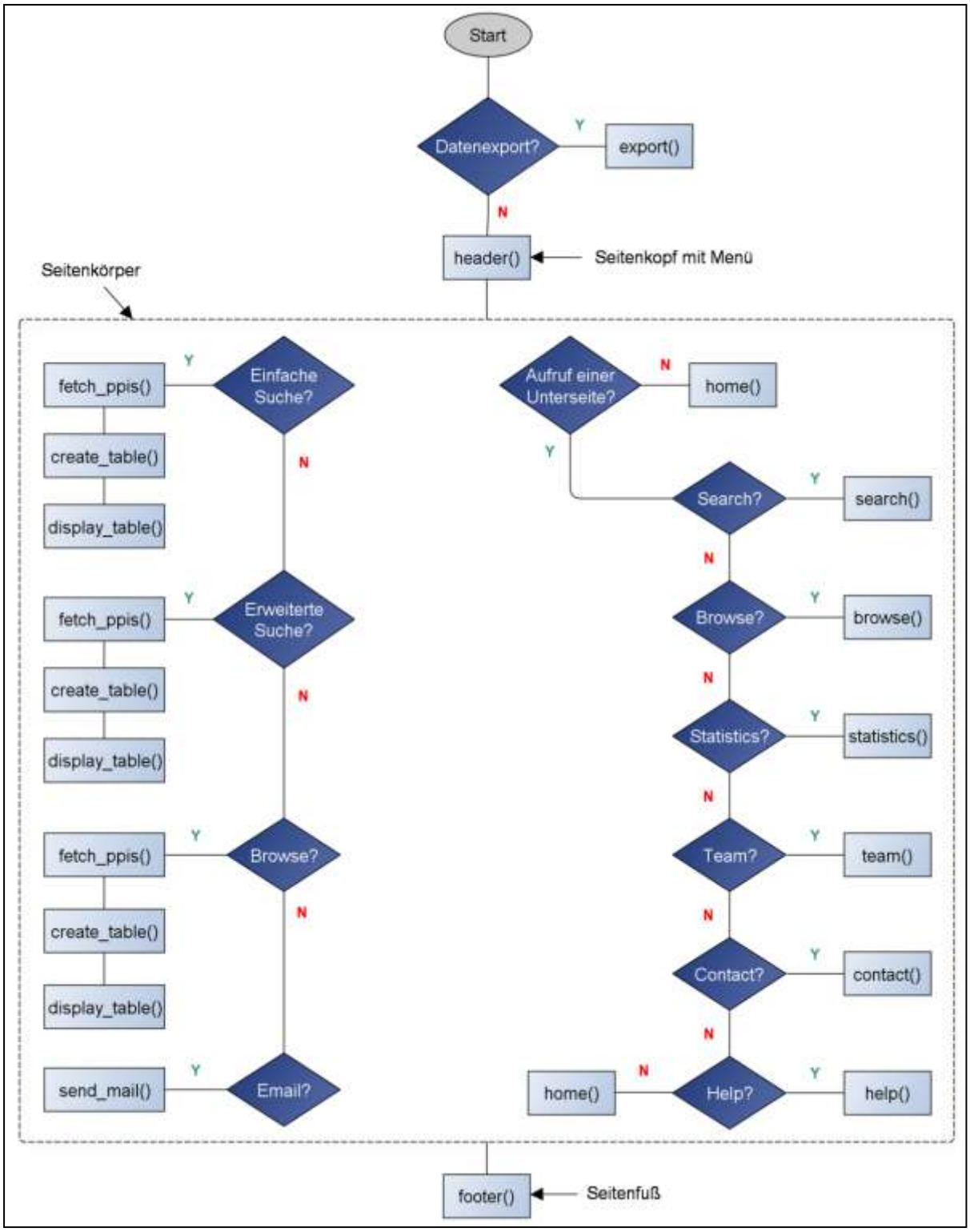


Abb. 9 Flussdiagramm des prozeduralen Aufrufs aller wichtigen Funktionen zur Erstellung der Webseite

Das Flussdiagramm verdeutlicht von oben nach unten den prozeduralen und kontrollierten Aufruf der wichtigsten Funktionen des Webseiten-Skriptes. Hellblaue Kästchen symbolisieren Funktionen, während dunkelblaue Rauten bedingte Verzweigungen darstellen, die eine Bedingung prüfen und je nach Ergebnis der Prüfung in eine Funktion oder eine andere Verzweigung führen. Z. B. bedeutet die erste im Skript auftretende Verzweigung *Datenexport?*, dass an dieser Stelle geprüft wird, ob eine Benutzereingabe zum Exportieren von Recherchedaten vorliegt. Trifft dies zu (Abzweigungen mit einem grünen Y für Yes), wird die Funktion *export()* zum Erstellen der entsprechenden Datei ausgeführt, welche einen von der Webseite losgelösten Abschnitt darstellt und daher ganz oben im Diagramm abgebildet ist. Andernfalls (Abzweigungen mit einem roten N für No)

werden nacheinander die Funktionen zur Erstellung der Webseite aufgerufen, angefangen mit der Funktion *header()*, welche den Kopfbereich der Webseite mit dem Navigationsmenü erstellt (s. Abb. 10). Danach wird die Funktion zur dynamischen Erstellung des mittleren Seitenkörpers aufgerufen (gestrichelt umrandeter Bereich). Diese Funktion prüft nacheinander, ob Benutzereingaben gemacht wurden und um welche es sich handelt, um dann die entsprechenden Funktionen aufzurufen. Im rechten Bereich des Diagramms wird überprüft, was im mittleren Seitenkörper dargestellt werden muss, also welcher Menüpunkt vom Benutzer angeklickt wurde. Im linken Teil wird überprüft, ob eine Benutzereingabe gemacht wurde. Zuletzt wird die Funktion *footer()* zur Erstellung des Seitenfußes aufgerufen.

4. ERGEBNISSE & DISKUSSION

4.1. Die PHISTO-Datenbank

4.1.1. Die Architektur der PHISTO-Datenbank

Die Datenbank besteht aus einer Tabelle mit den Kerndaten der PPI und drei Tabellen mit den beschreibenden Metadaten. Dank der Auftrennung der Daten in Kern- und Metadaten liegt die Datenbank in normalisierter Form vor und ist somit frei von jeglichen Redundanzen und zusätzlich einfach zu pflegen. Zudem machen die Verwendung von international gängigen Kennungen sowie der Einbezug von externen, wissenschaftlichen Datenbanken die gespeicherten PPI-Daten einheitlich und konsistent. Das auf Kennungen basierte Datenschema ermöglicht außerdem eine einfache Erweiterung der Datenbank und sichert dabei die Konsistenzerhaltung innerhalb der Datenbank.

4.1.2. Der Inhalt der PHISTO-Datenbank

In Tabelle 3 sind die wichtigsten statistischen Angaben zu den in der PHISTO-Datenbank gespeicherten Daten zu sehen. In Tabelle 4, 5 und 6 sind detaillierte Auflistungen zu den vier in der Datenbank gespeicherten pathogenen Obergruppen der Viren, Bakterien, Pilze und Protozoen zu sehen. Es sei darauf hingewiesen, dass die Statistiken am 23. Juni 2013 erstellt wurden und sich mit einer Aktualisierung der Datenbank ändern können.

Pathogen-Gruppe	Anzahl der Arten bzw. Stämme	Anzahl der PPI	Anzahl interagierender Proteine von Pathogenen	Anzahl interagierender humaner Proteine
Viren	247	14511	927	3973
Bacteria	46	9132	2654	5083
Mycota (Fungi)	3	5	4	5
Protozoa	5	15	8	15
Insgesamt	300	23661	3593	9075

Tab. 3 **Übersicht über die Mengen der in der Datenbank gespeicherten Pathogene, der Interaktionen und involvierten Proteine**

PPI zwischen humanen und viralen sowie bakteriellen Proteinen bilden den Großteil der in der Datenbank gespeicherten PPI. Dies liegt darin begründet, dass bis dato in größeren Studien zur Aufklärung interspezifischer Interaktome ausschließlich Viren und Bakterien betrachtet wurden. Detaillierte Übersichten mit Daten zu den einzelnen Pathogen-Familien sind in den Tabellen 4, 5 und 6 zu finden.

Viren

Familie	Anzahl der Arten	Anzahl der PPI	Anzahl interagierender Proteine von Viren	Anzahl interagierender humaner Proteine
Adenoviridae	12	125	33	77
Arenaviridae	5	16	5	13
Arteriviridae	1	56	1	56
Asfarviridae	1	3	3	2
Baculoviridae	1	1	1	1
Bunyaviridae	7	174	7	138
Circoviridae	2	4	2	3
Coronaviridae	2	7	6	6
Filoviridae	3	116	5	101
Flaviviridae	21	1058	173	424
Hepadnaviridae	8	45	16	42
Hepeviridae	2	3	2	2
Herpesviridae	31	1061	171	599
Myoviridae	3	4	3	3
Orthomyxoviridae	21	828	47	346
Papillomaviridae	15	333	58	135
Paramyxoviridae	11	856	22	638
Parvoviridae	1	7	5	4
Picornaviridae	6	9	8	5
Polyomaviridae	5	75	11	54
Poxviridae	11	215	36	150
Reoviridae	7	61	9	55
Retroviridae (davon 49 Unterarten des HIV)	62	9439	292	1108
Rhabdoviridae	4	9	5	6
Siphoviridae	2	2	2	2
Togaviridae	1	1	1	1
Nicht zuordnungsfähig	2	2	2	1

Tab. 4 **Detaillierte Übersicht über die in der Datenbank gespeicherten viralen Erreger, die von ihnen ausgehenden Interaktionen und die involvierten Proteine**

Die viralen Erreger machen mit insgesamt 247 Arten den weitaus größten Teil der in der Datenbank gespeicherten Pathogene aus. Das liegt daran, dass sich die noch recht junge Interaktomforschung bisher größtenteils mit den Interaktomen von Viren befasst hat. Mit 9439 Einträgen Interaktionen geht fast die Hälfte der gespeicherten Interaktionen auf die Familie der Retroviridae und innerhalb dieser auf HI-Viren zurück, die hier mit 40 Unterarten vertreten sind. Die hohe Anzahl an gespeicherten Interaktionen zu HI-Viren ist darauf zurückzuführen, dass dieses äußerst gefährliche und weit verbreitete Virus eines der am besten erforschten humanpathogenen Viren ist.

Bacteria

Familie	Anzahl der Arten bzw. Stämme	Anzahl der PPI	Anzahl interagierender Proteine von Bakterien	Anzahl interagierender humaner Proteine
Aeromonadaceae	1	2	1	2
Bacillaceae	2	3182	943	1751
Bacillus anthracis	1	3181	942	1750
Bacillus subtilis	1	1	1	1
Campylobacteraceae	1	3	1	3
Chlamydiaceae	2	21	3	21
Clostridiaceae	2	45	10	6
Enterobacteriaceae	14	4454	1303	2244
Citrobacter rodentium	1	1	1	1
Escherichia coli	5	40	25	35
Klebsiella pneumoniae	1	1	1	1
Salmonella enterica	2	7	7	7
Shigella flexneri	1	20	9	13
Yersinia enterocolitica	1	7	3	7
Yersinia pestis	2	4377	1260	2187
Yersinia pseudotuberculosis	1	1	1	1
Francisellaceae	2	1341	2	988
Francisellaceae tularensis	2	1341	2	988
Helicobacteraceae	3	5	4	5
Legionellaceae	1	1	1	1
Listeriaceae	2	7	6	5
Moraxellaceae	1	1	1	1
Mycoplasmatacae	1	1	1	1
Neisseriaceae	1	17	2	17
Peptostreptococcaceae	2	7	2	5
Pseudomonadaceae	1	14	3	10
Staphylococcaceae	4	16	13	14
Streptococcaceae	4	12	11	7

Tab. 5 **Detaillierte Übersicht über die Menge der in der Datenbank gespeicherten bakteriellen Erreger, die von ihnen ausgehenden Interaktionen und die involvierten Proteine**

Bei den Bakterienfamilien Bacillaceae, Enterobacteriaceae und Francisellaceae sind auch die beinhalteten Arten aufgelistet. Dadurch wird erkennbar, dass der Großteil der gespeicherten Interaktionen bei den Bakterien nur drei Bakteriengattungen zuzuordnen ist: *Bacillus anthracis*, *Yersinia pestis* und *Francisella tularensis*. Dies liegt daran, dass diese drei Bakterien die ersten und bisher einzigen sind, deren Interaktome in großangelegten Studien systematisch erforscht wurden.

Mycota (Fungi) und Protozoa

Familie	Anzahl der Stämme	Anzahl der PPI	Anzahl interagierender Proteine von Pathogenen	Anzahl interagierender humaner Proteine
Mycota (Fungi)				
Pneumocystidaceae	1	1	1	1
Radiomycetaceae	1	2	1	2
Nicht zuordnungsfähig	1	2	2	2
Protozoa				
Trypanosomatidae	1	1	1	1
Nicht zuordnungsfähig	4	14	7	14
Acanthamoeba castellanii	1	1	1	1
Plasmodium falciparum	1	7	5	7
Plasmodium vivax	1	1	1	1
Plasmodium yoelii	1	5	1	5

Tab. 6 **Detaillierte Übersicht über die in der Datenbank gespeicherten Pilze und Protozoen, die von ihnen ausgehenden Interaktionen und die involvierten Proteine**

Mycota und Protozoa sind im Vergleich zu Viren und Bakterien in der Datenbank viel seltener vertreten, da die pathogenen Gruppen bisher nicht das Ziel größerer Interaktom-Studeien gewesen sind. Dabei fällt insbesondere die geringe Anzahl an Interaktionen von Plasmodien auf, da diese Einzeller für die Tropenkrankheit Malaria verantwortlich ist, eine der aktuell häufigsten und gefährlichsten Infektionskrankheiten weltweit. Diese Tatsache lässt aber auch vermuten, dass in den kommenden Jahren Studien und somit mehr Interaktionsdaten zu diesem Erreger folgen werden.

Insgesamt umfasst die Datenbank 23.663 verschiedene experimentell nachgewiesene PPI zwischen 9.075 humanen Proteinen und 3.593 Proteinen von 300 verschiedenen pathogenen Organismen. Der Großteil der Interaktionen verteilt sich auf 247 virale (s. Tab. 4) und 46 bakterielle Vertreter (s. Tab. 5). Pilze sind insgesamt mit nur 3 Arten und Protozoen mit nur 5 Arten vertreten (s. Tab. 6).

Die sehr viel höhere Anzahl an PPI von Bakterien und insbesondere Viren gegenüber Pilzen und Protozoen liegt darin begründet, dass bisher alle großangelegten Studien zur Aufklärung von interspezifischen Interaktomen ausschließlich Viren und Bakterien als Forschungsobjekt hatten (Durmus Tekir S., 2013, S. 40). Bei den Viren dominieren die HI-Viren mit mehr als 9000 (etwa 40% aller gespeicherten Interaktionen) Einträgen die gesamte Datenbank. Dies liegt darin begründet, dass AIDS als eine der gefährlichsten und am weitesten verbreiteten Infektionskrankheiten und als häufigste infektionsbedingte Todesursache in der dritten Welt

auch das am besten erforschte Virus ist. Mit einem äußerst kleinen Genom und Proteom ist das HI-Virus wie andere RNA-Viren in hohem Maße vom zellulären Vermehrungsapparat abhängig. Die viralen Interaktionen zu den Molekülen des zellulären Vermehrungsapparates ermöglichen dem HI-Virus die Replikation und gelten daher als molekularbiologische Schlüsselmechanismen für ein vollständiges Verständnis der Pathogenese von HI-Viren (Durmus Tekir S., 2013, S. 17). Somit bot sich das HI-Virus als Forschungsobjekt für Interaktomstudien geradezu an. Bei den Bakterien sind als erstes *Bacillus anthracis* (Erreger des Milzbrandes), *Yersinia pestis* (Erreger der Lungen- und Beulenpest) und *Francisella tularensis* (Erreger der Tularämie) intensiv und systematisch auf ihre interspezifischen Interaktome erforscht worden, weil diese drei hochansteckenden und tödlichen Pathogene zu den Mikroorganismen gehören, die bei einem potentiellen Missbrauch als biologische Waffe eine große Bedrohung für die Zivilbevölkerung darstellen (Dyer et al., 2010). Daher existieren innerhalb der in der Datenbank gespeicherten bakteriellen Interaktionen fast ausschließlich Daten zu diesen drei Erregern.

Der Fokus auf Viren und Bakterien rührt sicherlich daher, dass diese beiden Gruppen für die aktuell häufigsten und gefährlichsten Infektionskrankheiten wie AIDS (HI-Virus), Masern (Measles-Virus), Hepatitis B (Hepatitis-B-Virus), Tuberkulose (*Mycobacterium tuberculosis*), Cholera (*Vibrio cholerae*) und weitere infektionsbedingte Durchfallerkrankungen, Tetanus (*Clostridium tetani*), Syphilis (*Treponema pallidum*) sowie etliche infektionsbedingte Atemwegserkrankungen verantwortlich sind. Zudem sind die Interaktome von Viren aufgrund ihrer viel geringeren Genom- und Proteomgröße als bei Prokaryoten und Eukaryoten schneller und einfacher zu erforschen. Nichtdestotrotz oder gerade deshalb fällt bei Betrachtung der Statistiken auf, dass zu den meisten der oben genannten humanpathogenen Erreger kaum oder keine Interaktionsdaten in der Datenbank existieren. Auch der einzellige Parasit Plasmodium ist mit insgesamt 13 Interaktionen eher spärlich vertreten, obwohl dieser für die Tropenkrankheit Malaria verantwortlich ist, welche zu den häufigsten sowie gefährlichsten Infektionskrankheiten weltweit zählt. Der Mangel von interspezifischen Interaktionsdaten zu solch wichtigen Pathogenen in der PHISTO-Datenbank bedeutet, dass diese Daten auch in renommierten primären Interaktionsdatenbanken wie IntAct oder Reactome nicht existieren, da diese als Datenquelle für die PHISTO-Datenbank dienen. Auch wenn diese Datenbanken als erste Anlaufstelle für neue Ergebnisse aus der Interaktionsforschung gelten, kann daraus aber nicht geschlussfolgert werden, dass schlichtweg keine Studien zur Aufklärung interspezifischer molekularer Interaktionen der oben genannten Pathogene existieren. Vielmehr verhält es sich so,

dass sehr viele Interaktionsdaten, die vor der Existenz dieser Datenbanken oder in kleinen Studien entdeckt wurden, in der entsprechenden naturwissenschaftlichen Literatur verborgen sind, und dass es bisher noch keiner Interaktionsdatenbank gelungen ist, all diese Daten zu finden und in die eigene Datenbank aufzunehmen. Beispielsweise enthält die PHISTO-Datenbank momentan nur 7 interspezifische PPI von Salmonellen, während in der Fachliteratur 38 weitere PPI aufgeführt sind (Schleker et al., 2012). Da die manuelle Suche nach in Frage kommender Literatur und die anschließende manuelle Suche nach Interaktionsdaten zu aufwendig ist, ist momentan die Entwicklung von neuarteigen Text Mining-Programmen von höchstem Interesse, welche mittels speziell für molekularbiologische Interaktionen entwickelter Algorithmen fortwährend im Internet nach relevanter Literatur suchen und diese automatisch nach Interaktionsdaten durchkämmen.

Letztlich sei bei der kritischen Betrachtung der in der PHISTO-Datenbank gespeicherten Daten und Datenmengen auch darauf hingewiesen, dass die Interaktomforschung noch in den Anfängen steht, doch die in den letzten Jahren kontinuierlich steigende Anzahl an Publikationen zu interspezifischen PPI und die rasanten Fortschritte in der Proteomforschung weisen darauf hin, dass in den nächsten Jahren etliche neue Interaktom-Studien und damit viele neue Interaktionsdaten zu bisher nicht oder kaum erforschten Pathogenen folgen werden. Diese Daten gilt es mittels automatisierter Verfahren von den primären Interaktionsdatenbanken zu beziehen und in die PHISTO-Datenbank einzupflegen, damit PHISTO auch zukünftig eine umfassende und aktuelle Plattform für PPI-Daten bleibt.

4.2. Die PHISTO-Webseite

Die Webseite als grafische Benutzeroberfläche zur anwenderseitigen, interaktiven Nutzung der PHISTO-Datenbank ist über die Internet-Adresse www.phisto.org erreichbar. Diese ist übersichtlich in einen obere Menüleiste und einen darunter befindlichen Darstellungsbereich geteilt (s. Abb. 10). Die Menüleiste ist wiederum mehrere Funktionalitäten bzw. Unterseiten gegliedert, zu denen man mit einem Klick auf das entsprechende Menüfeld gelangt. Der Darstellungsbereich dient der Anzeige der Funktionalitäten und Unterseiten, wenn diese im Menü angeklickt werden, sowie der Präsentation recherchierter Daten. Die wichtigsten Funktionalitäten zur Recherche in der PHISTO-Datenbank und die Datenpräsentation werden nachfolgend näher erläutert.

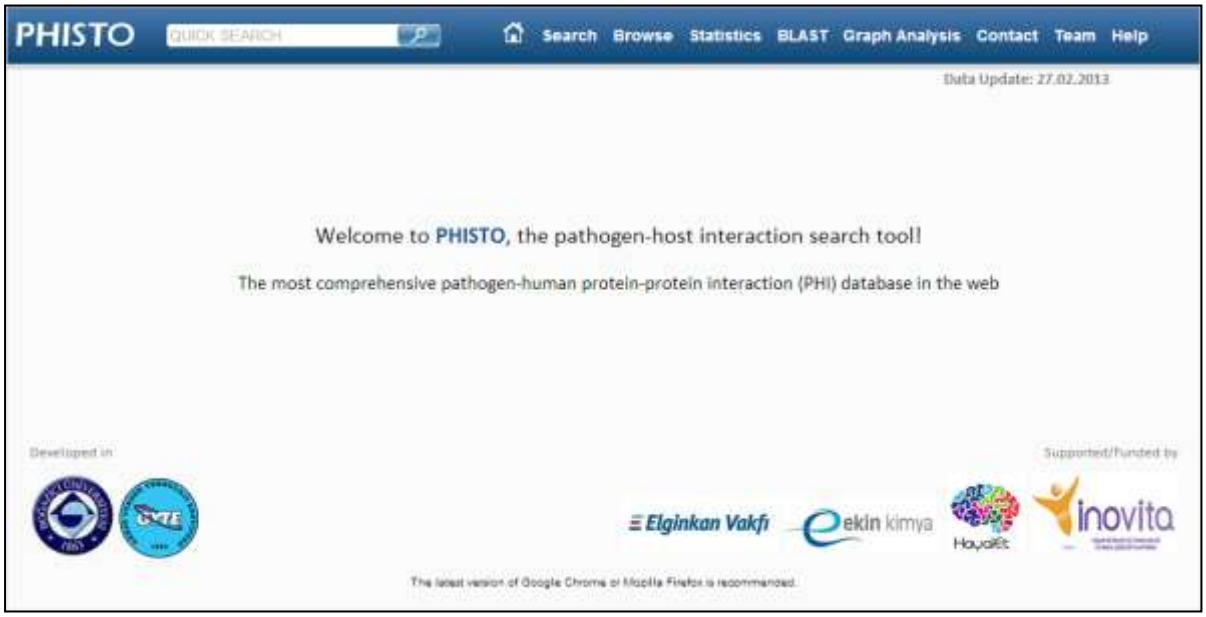


Abb. 10 **Die grafische Benutzeroberfläche von PHISTO auf der Webseite www.phisto.org**

Die Webseite ist in eine obere Menüleiste und einen Darstellungsbereich gegliedert. Die Menüleiste dient zur Navigation durch die verschiedenen Funktionalitäten für die PHISTO-Datenbank. Der Darstellungsbereich dient zur Anzeige der Funktionalitäten und zur Präsentation von Recherche-Ergebnissen. Die wichtigsten Funktionalitäten zur Datenrecherche werden in den nachfolgenden Unterkapiteln und Abbildungen näher erläutert.

4.2.1. Die einfache PPI-Recherche mit einem einzelnen Suchausdruck

Für die einfache Suche nach PPI-Daten anhand eines einzelnen Suchausdrucks können Anwender entweder die in die Menüleiste integrierte Schnell-Suche (s. Abb. 11a) oder die reguläre Such-Funktionalität verwenden, die über den Menüpunkt *Search* zu erreichen ist (s. Abb.

11b). Um eine Recherche über die Menüleiste oder die Such-Funktionalität zu starten, genügt das Drücken der Enter-Taste im Suchfeld. Alternativ können die entsprechenden Knöpfe angeklickt werden.

Die einfache Suche ist für schnelle PPI-Recherchen ohne großen Aufwand gedacht, bei denen die gesamten Informationen (ID-Nrn. der interagierenden und Bezeichnungen der interagierenden Proteine, Bezeichnung des Pathogens, Bezeichnung der experimentellen Nachweismethode, etc.) zu allen PPI in der Datenbank nach einem vom Benutzer angegebenen Ausdruck durchsucht werden. Dabei handelt es sich um eine unscharfe Suche, welche die Groß- und Kleinschreibung des Suchausdrucks nicht beachtet und prüft, ob der Ausdruck in einem Datenwert enthalten ist und nicht, ob der Ausdruck einem Datenwert enspricht.

Beispielsweise liefert die Suche nach dem Ausdruck „protease“ alle PPI, in denen die Bezeichnung des Proteins pathogenen Ursprungs oder die des humanen Proteins diesen Ausdruck enthält, also z. B. unter anderem eine PPI, in dem ein interagierendes Protein die Bezeichnung „Minor extracellular protease Vpr“ trägt.

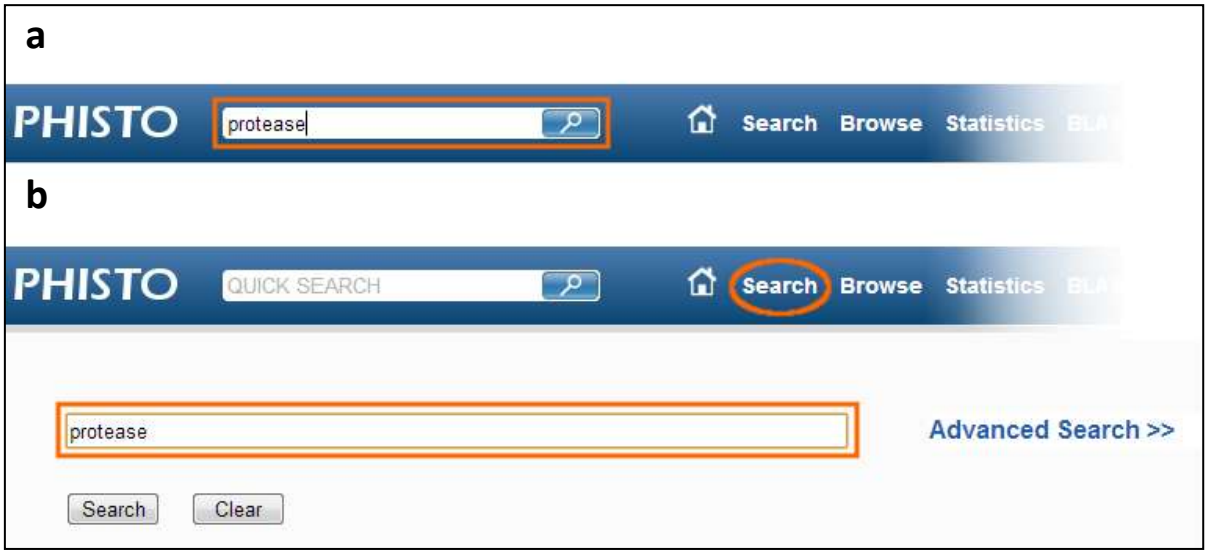


Abb. 11 Die einfache PPI-Recherche in der PHISTO-Datenbank

Eine einfache Suche kann entweder direkt in der Menüleiste (a) oder auf der Seite zur Datensuche (b) durchgeführt werden, die über einen Klick auf das Menüfeld *Search* erreichbar ist. Bei der einfachen Datensuche handelt es sich um eine ungenaue Suche des eingegebenen Ausdrucks in den gesamten Informationen aller PPI in der Datenbank. Hier ist beispielhaft die Suche nach dem Ausdruck „protease“ abgebildet. Die Ergebnisdarstellung der Suchanfrage ist in Abb. 13 zu sehen.

4.2.2. Die erweiterte PPI-Recherche mit mehreren Suchkriterien

Anders als die einfache Suche erlaubt die erweiterte Suche die Recherche nach PPI-Daten anhand mehrerer Suchkriterien. Die erweiterte Suche ist wie die einfache Suche über das Menüfeld *Search* zu erreichen, nur muss diese neben dem Eingabefeld zunächst aktiviert werden (s. Abb. 12). Durch das Klicken auf *Advanced Search* wird unterhalb des Suchfeldes ein neues Formular geöffnet, in dem beliebige Suchkriterien für eine möglichst exakte Datenrecherche festgelegt werden können. Neue Suchkriterien können durch das Anklicken von *Add another field* hinzugefügt werden.

Eine einzelnes Suchkriterium wird wie folgt erstellt: Zunächst wird in der ganz rechten Auswahlliste ein PPI-Datenfeld ausgewählt, in dessen Werten nach einem Ausdruck gesucht werden soll. Die Liste enthält eine voreingestellten Auswahl aus der taxonomischen ID-Nr. des Pathogens (*Taxonomy ID*), der Bezeichnung des Pathogens (*Pathogen*), der Uniprot-ID-Nr. des interagierenden Proteins pathogenen Ursprungs (*Pathogen Protein Uniprot ID*), der Bezeichnung des interagierenden Proteins pathogenen Ursprungs (*Pathogen Protein*), der Uniprot-ID-Nr. des interagierenden Proteins pathogenen Ursprungs (*Pathogen Protein Uniprot ID*), der Bezeichnung des interagierenden Proteins pathogenen Ursprungs (*Pathogen Protein*), der Bezeichnung der Nachweismethode (*Observation Method*) und die PubMed-ID-Nr. (*Pubmed ID*). Nach der Auswahl eines Datenfeldes wird in das danebengelegene Eingabefeld der Ausdruck eingegeben, nach welchem in dem Datenfeld gesucht werden soll. Wie bei der einfachen Suche handelt es sich hierbei um eine ungenaue Suche, sodass die Groß- und Kleinschreibung nicht beachtet werden muss. In der Auswahlliste links vom Eingabefeld kann eine logische Verneinung des Suchkriteriums eingestellt werden, indem der Wert *NOT* ausgewählt wird. Das bedeutet, dass der eingegebene Suchausdruck in dem angegebenen Datenfeld ausgeschlossen werden soll. Mit den beiden Auswahllisten ganz links können mehrere Suchkriterien miteinander verknüpft werden. Zum einen muss ausgewählt werden, ob das jeweilige Suchkriterium mit den vorigen Kriterien durch eine logische Konjunktion (*AND*) oder eine logische Disjunktion (*OR*) verknüpft werden soll, und zum anderem muss die Verknüpfungsebene des Suchkriteriums festgelegt werden. Sind auf diese Weise alle gewünschten Suchkriterien und die Verknüpfungen dieser Kriterien angegeben, kann die Recherche über Knopf *Search* gestartet werden.

Die Anwendung der erweiterten Suche soll anhand des in Abb. 12 verwendeten Beispiels erläutert werden: Es soll nach allen Interaktionen gesucht werden, die von *E. Coli*-Proteinen ausgehen und die entweder durch das Hefe-Zwei-Hybrid-System oder durch ein Pulldown-Assay nachgewiesen wurden, ausgenommen die Interaktionen von Proteinen des *E. Coli*-Stammes K-12. Dazu wird als erstes Kriterium angegeben, dass in dem Feld *Pathogen* nach dem Ausdruck „Escherichia“ gesucht werden soll, und gleich im zweiten Kriterium, dass im selben Feld der Ausdruck „K-12“ ausgeschlossen werden soll. Das dritte und vierte Kriterium geben an, dass in dem Feld *Observation Method* nach den beiden Ausdrücken „two hybrid“ und „pull down“ gesucht werden soll. Diese beiden Kriterien werden durch ein logisches Oder verknüpft und auf eine gemeinsame Verknüpfungsebene gestellt, sodass wie bei einer Klammerung zuerst diese beiden Kriterien überprüft werden.

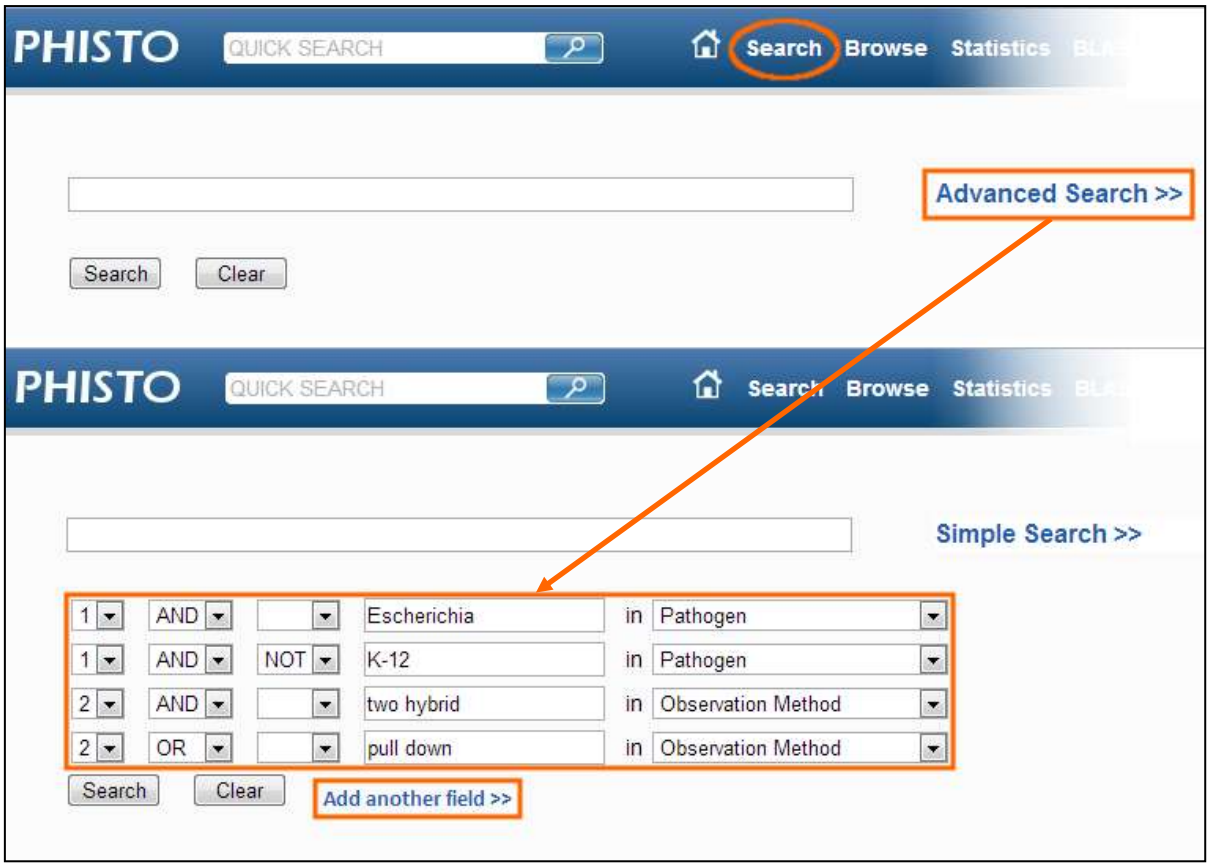


Abb. 12 **Die erweiterte PPI-Recherche in der PHISTO-Datenbank**

Eine erweiterte Recherche kann auf der Seite *Search* gemacht werden, indem auf *Advanced Search* geklickt wird. Dadurch öffnet sich ein neues Formular, indem zeilenweise Kriterien für die Datensuche angegeben werden können. Eine neue Bedingung kann hinzugefügt werden, indem auf *Add another field* geklickt wird. Eine einzelne Bedingung setzt sich von rechts nach links zusammen aus dem zu durchsuchenden Attribut, dem zu suchenden Ausdruck, einer eventuellen logischen Verneinung, der logischen Verknüpfungsart des Suchkriteriums und der Verschachtelungsebene. Die Ergebnisdarstellung der Suchanfrage ist in Abb. 13 zu sehen.

4.2.3. Die PPI-Recherche nach der taxonomischen Klassifikation der Pathogene

Neben den zuvor erläuterten Funktionalitäten zur Datenrecherche anhand von Suchausdrücken wird den Benutzern von PHISTO über den Menüpunkt *Browse* die Möglichkeit geboten, die in der Datenbank gespeicherten PPI-Daten nach der taxonomischen Einteilung der Pathogene zu durchsuchen (s. Abb. 13). Auf diese Weise kann einerseits die gesamte Datenbank ohne die Angabe von Suchkriterien „durchstöbert" werden und andererseits ein Überblick zu den pathogenen Gruppen verschafft werden, die in der PHISTO-Datenbank enthalten sind. Da in der Datenbank zu jedem pathogenen Organismus das Reich bzw. die Domäne, die Familie, die Spezies und ggf. der Stamm gespeichert ist, wird zu jeder dieser taxonomischen Klassifikationskategorien eine eigene Auswahlliste angezeigt, angefangen mit der übergeordneten Liste der Reiche bzw. Domänen, welche in PHISTO Viren, Bakterien, Protozoen und Pilzen beinhaltet. Nach der Auswahl einer Gruppe aus dieser Liste wie beispielsweise das Reich der Bakterien wird darunter die Auswahlliste für die nächsttiefere Klassifikationskategorie angezeigt, also alle Bakterienfamilien, die in der Datenbank existieren. In dieser Liste kann eine beliebige Familie wie z. B. die der Enterobacteriaceae gewählt werden, um eine Auswahlliste mit den zugehörigen Spezies anzuzeigen. Wird eine Spezies wie z. B. *Salmonella enterica*

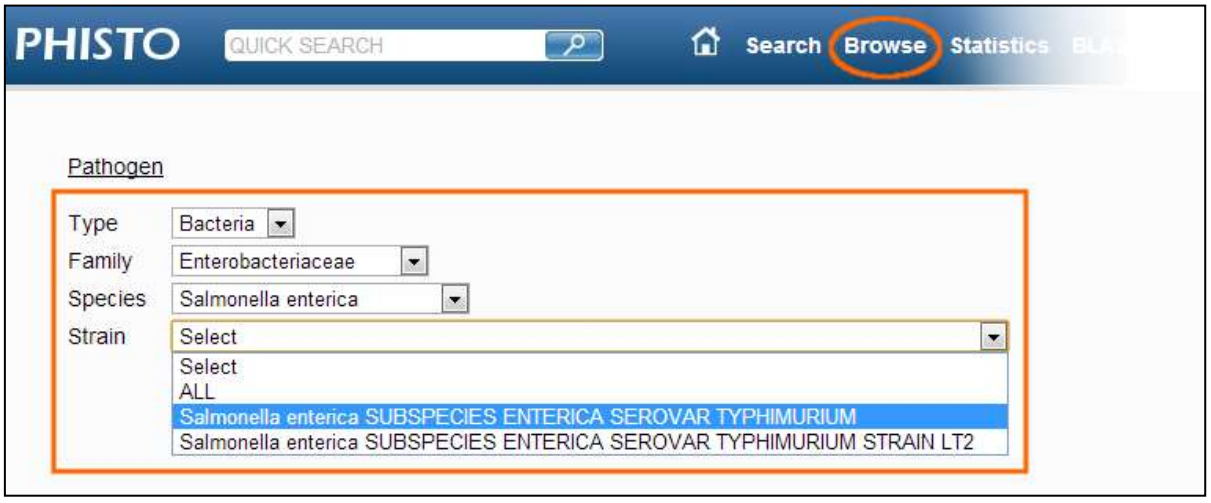


Abb. 13 Die PPI-Recherche nach der taxonomischen Klassifizierung der Pathogene

Auf der Seite *Browse* können alle in der Datenbank gespeicherten PPI gemäß der taxonomischen Klassifizierung aller Pathogene durchstöbert werden. Dazu dienen vier Auswahllisten zu jeder Klassifizierungskategorie, in die jedes einzelne Pathogen zugeordnet ist. Als erstes ist eine Domäne/ein Reich zu wählen (hier beispielhaft *Bacteria*). Innerhalb dieser Gruppierung ist eine Familie zu wählen (hier beispielhaft *Enterobacteriaceae*). Innerhalb dieser Familie ist eine Spezies zu wählen (hier beispielhaft *Salmonella enterica*). Innerhalb dieser Spezies ist ein Stamm zu wählen (*hier beispielhaft Salmonella enterica SUBSPECIES ENTERICA SEROVAR TYPHIMURIUM*). Optional kann in jeder Liste der Wert *ALL* ausgewählt werden, sodass nach den PPI aller Pathogene der jeweiligen Klassifikationskategorie gesucht wird. Die Ergebnisdarstellung ist in Abb. 13 zu sehen.

gewählt, erscheint die letzte Auswahlliste mit eventuellen Stämmen zu dieser Spezies. In jeder Auswahlliste wird zusätzlich der Wert *All* angeboten, um nach den PPI aller pathogenen Organismen der jeweiligen Klassifikationskategorie zu suchen. So werden beispielsweise die PPI aller Bakterien gesucht, wenn nach in der Liste für die Bakterienfamilien *All* ausgewählt wird.

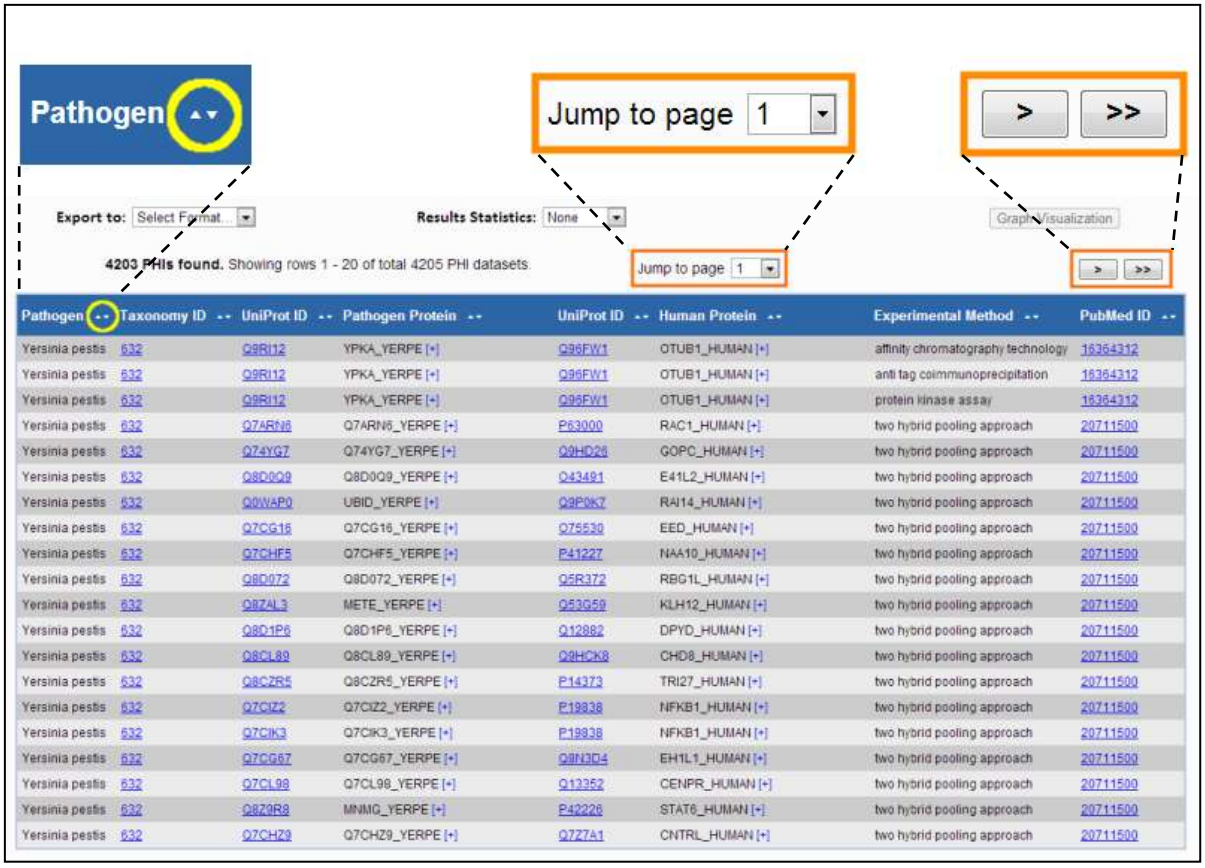
Momenten beschränkt sich die *Browse*-Funktionalität auf die taxonomische Klassifikation der Pathogene, doch es könnte sinnvoll sein, die in der Datenbank gespeicherten PPI nach anderen Klassifikationen durchsuchen zu können, z. B. nach Gruppierungen der Proteine. So könnte als erste Klassifikationskategorie die Funktion des Proteins angegeben werden: Enzym, Strukturprotein, Antikörper, Toxin, Rezeptor, Hormon, etc. Die Einteilung könnte beispielsweise bei den Enzymen durch die international gängige Klassifikation nach den von ihnen katalysierten Reaktionen vertieft werden: Die erste Klassifikationsebene wäre die Einteilung in Oxidoreduktasen, Transferasen, Hydrolasen, Lyasen, Isomerasen sowie Ligasen und die zweite beispielhaft für die Oxidoreduktasen in Dehydrogenasen, Hydrogenasen, Oxidasen, Oxygenasen sowie Hydroxylasen. Die Erweiterung um solch eine weitere Klassifikation ist sowohl datenbankseitig als auch in der Webseite schnell und unkompliziert umsetzbar. In der Datenbank müsste die Tabelle *Proteins* (s. Abb. 7) um drei oder vier Spalten für die Klassifikationen der Proteine erweitert werden. Die entsprechenden Datenwerte könnten für jedes Protein dank der UniProt-ID-Nr. von der Datenbank UniProt bezogen werden. Für die Erweiterung in der Webseite könnte dieselbe Funktion verwendet werden, die auch für die taxonomische Klassifikation verwendet wird, nur dass andere Parameter verwendet werden müssten.

4.2.4. Die Präsentation von Rechercheergebnissen

Die Ergebnisse aller Recherchen werden in einem übersichtlichen, tabellarischen Format mit festgelegten Spalten dargestellt, deren Namen in einem blauen Tabellenkopf dargestellt sind (s. Abb. 14). Von links nach rechts besteht die Tabelle aus folgenden Spalten: *Pathogen* für die Bezeichnung des Pathogens, *Taxonomie ID* für die Taxonomie-Kennung des Pathogens, *UniProt ID* für die UniProt-Kennung des Proteins des Pathogens, *Pathogen Protein* für die Bezeichnung des Proteins des Pathogens, *UniProt ID* für die Uniprot-Kennung des humanen Proteins, *Human Protein* für die Bezeichnung des humanen Proteins, *Experimental Method*

für die Bezeichnung der experimentellen Nachweismethode, PubMed ID für die PubMed-Kennung der Publikation. Neben jedem Spaltennamen ist ein nach oben zeigendes und ein nach unten zeigendes Dreieck abgebildet. Ein Klick auf eines der Dreiecke ermöglicht die auf- oder absteigende, alphabetische und numerische Sortierung der Tabelle nach der entsprechenden Spalte.

Unterhalb des Tabellenkopfes werden zeilenweise die PPI-Datensätze aufgelistet, welche die Recherche ergeben hat. Die Kennungen innerhalb eines Datensatzes verweisen alle auf externe Webseiten anderer Datenbanken, um Benutzern die Möglichkeit zu bieten, mehr Informationen zu dem jeweiligen Objekt einzuholen. Erkennbar sind diese an der blauen Schriftfarbe,



Pathogen	Taxonomy ID	UniProt ID	Pathogen Protein	UniProt ID	Human Protein	Experimental Method	PubMed ID
Yersinia pestis	632	Q9RI12	YPKA_YERPE [+]	Q96FW1	OTUB1_HUMAN [+]	affinity chromatography technology	16364312
Yersinia pestis	632	Q9RI12	YPKA_YERPE [+]	Q96FW1	OTUB1_HUMAN [+]	anti tag coimmunoprecipitation	16364312
Yersinia pestis	632	Q9RI12	YPKA_YERPE [+]	Q96FW1	OTUB1_HUMAN [+]	protein kinase assay	16364312
Yersinia pestis	632	Q7ARN6	Q7ARN6_YERPE [+]	P63000	RAC1_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q74YG7	Q74YG7_YERPE [+]	Q9HD26	GOPC_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q8D0Q9	Q8D0Q9_YERPE [+]	O43491	E41L2_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q0WAP0	UBID_YERPE [+]	Q9P0K7	RAI14_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q7CG16	Q7CG16_YERPE [+]	O75530	EED_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q7CHF5	Q7CHF5_YERPE [+]	P41227	NAA10_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q8D072	Q8D072_YERPE [+]	Q5R372	RBG1L_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q8ZAL3	METE_YERPE [+]	Q53G59	KLH12_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q8D1P6	Q8D1P6_YERPE [+]	Q12882	DPYD_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q8CL89	Q8CL89_YERPE [+]	Q9HCK8	CHD8_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q8CZR5	Q8CZR5_YERPE [+]	P14373	TRI27_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q7CIZ2	Q7CIZ2_YERPE [+]	P19838	NFKB1_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q7CIK3	Q7CIK3_YERPE [+]	P19838	NFKB1_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q7CG67	Q7CG67_YERPE [+]	Q8N3D4	EH1L1_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q7CL98	Q7CL98_YERPE [+]	Q13352	CENPR_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q8Z9R8	MNMG_YERPE [+]	P42226	STAT6_HUMAN [+]	two hybrid pooling approach	20711500
Yersinia pestis	632	Q7CHZ9	Q7CHZ9_YERPE [+]	Q7Z7A1	CNTRL_HUMAN [+]	two hybrid pooling approach	20711500

Abb. 14 Die Ergebnisdarstellung von PPI-Recherchen in Tabellenform

Die Ergebnisse von PPI-Recherchen werden in Form einer Tabelle mit festgelegten Spalten dargestellt. Die Tabelle ist in einen Tabellenkopf mit den Spaltennamen (blauer Balken) und einen darunterliegenden Tabellenkörper aufgeteilt, in dem zeilenweise die PPI-Datensätze aufgelistet werden. Im Tabellenkopf sind neben jedem Spaltennamen Symbole zur alphabetischen und numerischen Sortierung der Tabelle nach der entsprechenden Spalte dargestellt (gelb umkreist). Zu externen Webseiten verweisende Hyperlinks sind in blauer Schrift dargestellt. Ein Klick auf den Link öffnet ein neues Fenster der entsprechenden Webseite, in dem weitere Informationen zu dem Datenelement bezogen werden können. In der Tabelle werden immer maximal 30 PPI angezeigt. Für die Navigation durch die weiteren PPI dienen verschiedene Bedienelemente oberhalb der Tabelle (orange umrandet).

welche in der Webseite einen Hyperlink, also einen elektronischen Querverweis auf eine andere Webseite, darstellen. Die Taxonomie-Kennungen verweisen auf Webseiten von NCBI, auf der nähere Informationen zum entsprechenden Pathogen nachgeschaut werden können. Die über die UniProt-Kennungen verwiesenen Webseiten von UniProt bieten umfangreiche Informationen zu den interagierenden Proteinen. Die PubMed-Kennungen verweisen auch auf Webseiten von NCBI. In diesen können Informationen zu der entsprechenden Publikation in Erfahrung gebracht werden.

In der Tabelle werden immer nur 30 PPI-Datensätze angezeigt, um die Seite übersichtlich zu halten. Um zu den anderen Datensätzen zu navigieren, kann entweder die *Jump to page*-Option oberhalb der Tabelle verwendet werden, mit der zu einer beliebigen Stelle der Tabelle gewechselt wird, oder die Knöpfe oberhalb der Tabelle, die mit Pfeil-Symbolen versehen sind. Diese navigieren zu den nächsten bzw. vorigen 30 Datensätzen oder an das Ende bzw. den Anfang der Tabelle.

Momentan ist die Ergebnistabelle auf die geschilderten 8 Spalten beschränkt. Auch wenn den Benutzern die Möglichkeit geboten wird, über Links zu externen Webseiten anderer Datenbanken weitere Informationen einzuholen, ist es sicherlich sinnvoll, auch die Möglichkeit anzubieten, die Ergebnistabelle um weitere Informationsspalten zu erweitern. Dazu müssten in der Datenbank die bereits existierenden Tabellen erweitert oder neue Tabellen angelegt werden und die gewünschten Daten wie in Kap. 3.2.1. erläutert aus externen Datenbanken bezogen werden. Die Erweiterung in der Webseite ist einfach zu realisieren, da die Funktion zur Erstellung und Anzeige der Tabelle dynamisch arbeitet.

4.2.5. Die Export recherchierter PPI-Daten

Alle recherchierten PPI-Daten können mittels einer Export-Funktionalität in verschiedene Dateiformate für tabellarisch strukturierte Daten exportiert und runtergeladen werden. Dazu dient eine Auswahlliste oberhalb der Ergebnistabelle (s. Abb. 15), die folgende Formate anbietet: TXT (einfache Textdatei), TAB (Textdatei, in der Spaltenwerte durch Tabulator-Einzüge getrennt sind), CSV (Textdatei, in der Spaltenwerte durch Kommata bzw. Semikola getrennt sind) und XLS (Microsoft Excel-Datei). Aus dieser muss ein Dateiformat gewählt werden, damit die Datei erstellt und zum Download angeboten wird.

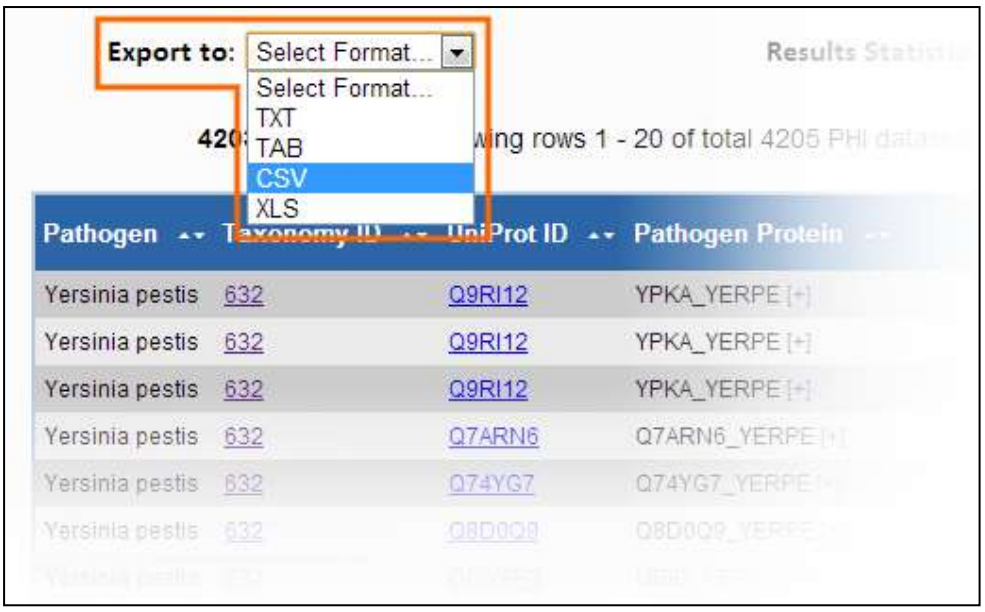


Abb. 15 Details der Ergebnisdarstellung von PPI-Recherchen: Optionen für den Export von PPI-Daten

Alle PPI-Daten der PHISTO-Datenbank können runtergeladen werden. Dazu müssen zunächst die gewünschten Daten in der Datenbank mittels einer der Suchfunktionalität angefragt werden (siehe Abb. 10, 11, 12). Anschließend kann in einer Auswahlliste oberhalb der Ergebnistabelle (orange markiert) das Dateiformat ausgewählt werden, in das die Daten gespeichert und zum Download angeboten werden sollen.

4.2.6. Weitere Features & Funktionalitäten

Neben den bereits erläuterten Funktionalitäten zur Datenrecherche, Ergebnispräsentation und zum Exportieren von Rechercheergebnissen bietet die Webseite für PHISTO weitere gängige Features und Funktionalitäten an, die über ein eigenes Menüfeld aufrubar und auf einer eigenen Seite angezeigt werden. Auf der Seite *Statistics* werden die wichtigsten Statistiken zu den Daten gezeigt, die in der Datenbank gespeichert sind. Auf der Seite *Help* können Anleitungen zur Bedienung der verschiedenen Funktionalitäten in der Webseite zu Hilfe genommen werden können. Auf der Seite *Contact* steht ein Kontaktformular zur Verfügung, in dem Benutzer Fragen, Kritiken und Anregungen an das Entwicklungsteam richten können. Auf der Seite *Team* sind alle Personen aufgelistet, die an der Entwicklung von PHISTO beteiligt waren. Desweiteren wurden von anderen Entwicklern die Funktionalitäten *BLAST* (Basic Local Alignment Search Tool) und *Graph Analysis* auf Basis der in dieser Bachelor-Arbeit entwickelten Datenbank und Webseite integriert. BLAST ist ein von Altschul et al. entwickelte Sammlung von Algorithmen und Programmen zur Analyse und zum Vergleich von biologischen Sequenzdaten (Altschul et al., 1990). Hier kann das BLAST-Interface verwendet werden, um die Aminosäuresequenz eines beliebigen Proteins einzugeben und diese mit den Se-

quenzen aller in der Datenbank gespeicherten Proteine vergleichen zu lassen, um homologe Interaktionsproteine zu finden (Durmus Tekir S., 2013, S. 35). Das *Graph Analysis*-Interface dient zur grafischen Darstellung verteilungsstatistischer Zusammenhänge innerhalb einer Gruppe von PPI (Durmus Tekir S., 2013, S. 36).

4.3. Die Verwendung der Datenbankanwendung PHISTO in der Forschung

Dank der in dieser Bachelorarbeit entwickelten Datenbank und der Webseite mit ihren verschiedenen Funktionen zur interaktiven Recherche sowie Analyse von PPI-Daten konnte Durmus Tekir die für ihre Doktorarbeit grundlegenden, bioinformatischen Datenanalysen zur Erforschung von allgemeinen Infektionsmechanismen von pathogenen Organismen durchführen (Durmus Tekir S., 2013, S. 40-84). In ihrer Arbeit verwendet sie die PPI-Daten, um die daraus resultierenden Interaktionsnetzwerke zwischen dem menschlichen Proteom und verschiedenen pathogenen Gruppen zu vergleichen.

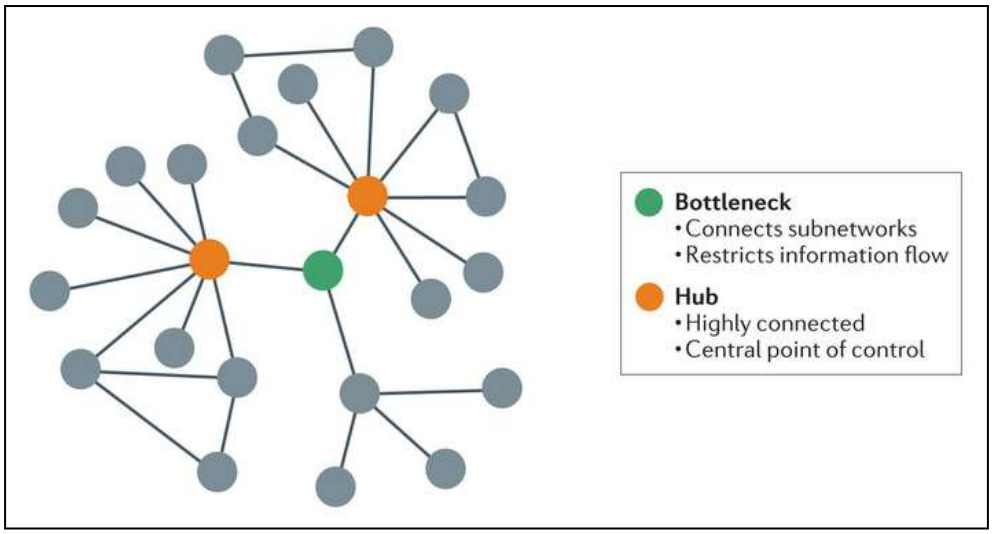


Abb. 16 Beispiel für Bottleneck- und Hub-Proteine innerhalb eines Proteinnetzwerkes

Zu sehen ist ein Beispiel für ein humanes Proteinnetzwerk. Sog. Bottlenecks (grüner Knoten) sind Proteine von geringer Konnektivität, die brückenartig kleinere Unternetzwerke verknüpfen und dadurch eine Art Engpass innerhalb des gesamten Netzwerkes bilden. Sog. Hubs sind dagegen stark vernetzt und fungieren dadurch als zentrale Kontrollknoten. Bottlenecks und Hubs sind häufig das Ziel von Infektionsmechanismen pathogener Organismen. Die Analyse der Interaktionen zu solchen Proteinen könnten wertvolle Erkenntnisse für die Entwicklung neuer Behandlungsmethoden und neuer Pharmaka liefern.

Vergl. Lynn et al., *Systems Virology: host-directed approaches to viral pathogenesis and drug targeting*, 2013

Dabei konnte unter anderem festgestellt werden, dass die Interaktion von Proteinen pathogener Organismen mit sog. Bottleneck-Proteinen und Hub-Proteinen humanen Ursprungs ein gängiger Infektionsmechanismus von allen Pathogenen ist (Durmus Tekir S., 2013, S. 85). Bottlenecks sind Proteine von geringer Konnektivität, die brückenartig Unternetzwerke des Gesamtnetzwerkes verknüpfen und dadurch eine Art Engpass bilden, während Hubs stark vernetzt sind und zentrale Kontrollknoten darstellen (s. Abb. 16). Desweiteren konnte Durmus Tekir Proteine identifizieren, die häufig und artenübergreifend von Proteinen pathogenen Ursprungs anvisiert werden, beispielsweise *P53*, ein Tumorsuppressor, der bei der Regulation des Zellzyklus mitwirkt. Die genauere Analyse der Interaktionen sowie der Funktionen der Interaktionen solcher Proteine könnten neue Erkenntnisse darüber liefern, wie Infektionen entstehen und vermieden werden können sowie zur Entwicklung neuartiger pharmakologischer Wirkstoffe beitragen.

Die Verwendung von PHISTO für solche Forschungen und die daraus resultierenden Erkenntnisse können sicherlich als Erfolg und Bestätigung dafür gewertet werden, dass mit dieser Bachelorarbeit eine wissenschaftlich relevante Plattform für die bioinformatische Interaktomforschung geschaffen wurde. Trotzdem und auch deswegen sollte PHISTO zukünftig weiterentwickelt werden, um durch die Integration spezifischerer Interaktionsdaten und die Implementierung weiterer analytischer Funktionalitäten eine noch umfassendere und zentralere Anlaufstelle für die Interaktomforschung zu bilden.

5. ZUSAMMENFASSUNG

Mit dieser Bachelorarbeit wurde die neue Datenbankanwendung PHISTO zur Recherche und Analyse von Protein-Protein-Interaktionen zwischen pathogenen Strukturen und dem menschlichen Organismus als Beitrag zur interspezifischen Interaktomforschung geschaffen. PHISTO beruht auf einer sekundären in MySQL entwickelten Datenbank, die ihre Daten aus neun externen Interaktionsdatenbanken und drei weiteren Datenbanken für die Metadaten bezieht. Dank der Vereinheitlichung der gespeicherten Daten gemäß eines internationalen Standards für Interaktionsdaten und ihrer logischen Auftrennung in mehrere Tabellen besitzt die Datenbank eine konsistente Architektur, die zudem einfach erweitert werden kann. Ein einheitliches Datenformat sowie die Verwendung international gängiger Kennungen und Bezeichnungen gewährleisten außerdem eine für Forschungsarbeiten optimale Weiterverwendung der Daten. Die als Benutzeroberfläche entwickelte dynamische Webseite ermöglicht Anwendern in übersichtlicher Form verschiedene Möglichkeiten der interaktiven Datenrecherche in dieser Datenbank. Mit der einfachen Datensuche kann die Datenbank schnell nach einem einzigen Ausdruck durchsucht werden. Bei der erweiterten Datensuche können dagegen mehrere, genaue Suchkriterien angegeben werden, um eine möglichst genaue Datenfilterung zu erreichen. Zusätzlich lässt sich die Datenbank bequem anhand der taxonomischen Klassifikation der gespeicherten Pathogene durchsuchen. Darüber hinaus konnten basierend auf der erweiterbaren Software-Plattform, die in dieser Arbeit entwickelt wurde, von anderen Entwicklern verschiedene Funktionalitäten zur grafischen Visualisierung und zur Analyse recherchierter Daten integriert werden.

Die Datenbank umfasst insgesamt mehr als 23.600 experimentell nachgewiesene interspezifische PPI von 300 verschiedenen pathogenen Organismen und ist damit die momentan umfangreichste, frei zugängliche Pathogen-Mensch-spezifische PPI-Datenbank. Der Großteil aller gespeicherten Interaktionen geht auf nur neun virale und bakterielle Pathogene zurück und spiegelt damit wider, dass bisher nur ein Bruchteil aller bekannten Pathogene umfassend hinsichtlich ihres Interaktoms erforscht worden sind. Allerdings deutet das enorme Potential der noch jungen Interaktomforschung, zum Verständnis von Infektionskrankheiten und zur Entwicklung neuer Therapieansätze sowie neuartiger pharmakologischer Wirkstoffe beizutragen, darauf hin, dass in Zukunft noch weitere Studien zur Aufklärung der Interaktome bisher nicht erforschter Pathogene folgen werden. Die daraus resultierenden Daten sollten mittels automatisierter Verfahren in die PHISTO-Datenbank eingepflegt werden, damit diese auch

zukünftig aktuell bleibt. Weitere Interaktionsdaten könnten mittels speziell entwickelter Text Mining-Algorithmen aus wissenschaftlichen Veröffentlichungen der molekularen Interaktionsforschung bezogen werden, die bisher von keiner Datenbank erfasst wurden. Dadurch könnte die Gesamtheit aller bisher nachgewiesener PPI erfasst und damit die erste vollständige Pathogen-Mensch-spezifische Interaktionsdatenbank geschaffen werden.

LITERATURVERZEICHNIS

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., et al. (2005). The Biomolecular Interaction Network Database and Related Tools: 2005 update. *Nucleic Acids Research*(33), S. D418-D424.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*, S. 403-441.

Balzert, H. (2007). *Basiswissen Web-Programmierung.* Herdecke/Witten: W3L GmbH.

Bünninge, M. (2011). *New Strategies to Fight Infectious Diseases - Arms Race on a Microscale.* infectionresearch.

Cannataro, M., & Guzzi, P. H. (2011). *Data Management of Protein Interaction Networks.* Hoboken, New Jersey, USA: Wiley.

Croft, D., O'Kelly, G., Wu, G., & Haw, R. (2011). Reactome: A Database of Reactions, Pathways and Biological Processes. *Nucleic Acids Research*(39), S. D691-D697.

Domling, A., Mannhold, R., Kubinyi, H., & Folkers, G. (2013). *Protein-Protein Interactions in Drug Discovery.* Weinheim: WILEY-VCH.

Durmus Tekir, S. (2013). *INVESTIGATION OF INFECTION MECHANISMS THROUGH PATHOGEN-HUMAN PROTEIN-PROTEIN INTERACTIONS BY BIOINFORMATICS APPROACHES.* Istanbul: Bogazici University.

Durmus Tekir, S., Cakir, T., & Ülgen, K. Ö. (2012). Infectious strategies of bactieral and viral pathogens through pathogen-human protein-protein-interactions. *Front. Microbio.*(3:46), S. 74-84.

Dyer, M. D., Neff., C., Dufford, M., Rivera, C. G., Shattuck, D., Bassaganya-Riera, J., et al. (2010). The Human-Bacterial Pathogen Interaction Networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. *PLOS ONE*(5:8), e12089.

Geisler, F. (2009). *Datenbanken: Grundlagen und Design.* Österreich: mitp-verlag.

Gillespi, J. J., Wattam, A. R., Cammer, S. A., Gabbard, J. L., Shukla, M. P., Dalay, O., et al. (2011). PATRIC: The Comprehensive Bacterial Bionformatics Resource with a Focus on Human Pathogenic Species. *Infection and Immunity*(79), S. 4286-4298.

Giralt, E., Peczuh, M., & Salvatella, X. (2011). *Protein Surface Recognition: Approaches for Drug Discovery.* Chichester: John Wiley & Sons.

Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H., Lamb, T., & Uetz, P. (2008). MPIDB: The Microbial Protein Interaction Database. *Bioinformatics*(24), S. 1743-1744.

Kemper, A. (2006). *Datenbanksysteme: Eine Einführung.* München: Oldenbourg Wissenschaftsverlag.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., et al. (2012). The IntAct Molecular Interaction Database in 2012. *Nucleic Acids Research*(40), S. D841-D846.

Kleinschmidt, P., & Rank, C. (2004). *Relationale Datenbanksysteme: Eine praktische Einführung.* Berlin Heidelberg: Springer.

Klussmann, E., & Scott, J. (2008). *Protein-Protein Interactions as New Drug Targets.* Heidelberg Berlin: Springer.

Kumar, R., & Nanduri, B. (2010). HPIDB - A Unifieed Resource for Host-Pahtogen Interactions. *BMC Bioinformatics*(11), S. 16.

Law, G. L., Korth, M. J., Benecke, A. G., Katze, M. G. (2013). Systems virology: host-directed approaches to viral pathogenesis and drug targeting. *Nature Review Microbiology*(11), S. 455-466.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeto, E., et al. (2012). MINT, the Molecular Interaction Database: 2012 update. *Nucleic Acids Research*(40), S. D857-D861.

Meier, A. (2004). *Relationale Datenbanken.* Berlin Heidelberg: Springer.

Nussinov, R., & Schreiber, G. (2009). *Computational Protein-Protein Interactions.* Boca Raton, Florida, USA: CRC Press.

Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stümpflen, V, et al. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology*(8), S. 894-898.

Prieto, C., & Rivas, J. D. (2006). APID: Agile Protein Interaction Data Analyzer. (34), S. W298-W302.

Razick, S., Magklara, G., & Donaldson, I. M. (2008). iRefIndex: A Consolidated Protein Interaction Database with Provenance. *BMC Bioinformatics*(9), S. 405.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*(32), S. D449-D451.

Schleker, S., Sun, J., Raghavan, B., Srnec, M., Muller, N., Koepfinger M., et al. (2012). The current Salmonella-host interactome. *Proteomics Clin. Appl.*(6), S. 117-133.

Srivastava, S. (2005). *Informatics In Proteomics.* Boca Raton, Florida, USA: CRC Press.

Stark, C., Breitkreutz, B. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., et al. (2011). The BioGRID Interaction Database: 2011 Update. *Nucleic Acids Research*(39), S. D698-D704.

Stebbins, C. E. (2005). Structural Microbiology at the Pathogen-Host Interface. *Cellular Microbiology, 7*, S. 1227-1236.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The STRING Database in 2011: Function Interaction Neworks of Proteins Globally Integrated and Scored. *Nucleic Acids Research*(39), S. D561-D568.

Walker, J. M., & Rapley, R. (2009). *Molecular Biology and Biotechnology.* Camridge, UK: RSCPublishing.

Winnenburg, R., Urban, M., Beacham, A., Baldwin, T. K., Holland, S., Lindeberg, M., et al. (2008). PHI-base Update: Additions to the Pathogen Host Interaction Database. *Nucleic Acids Research*, S. D572-D576.

World Health Organisation. (2012). *Global Report for Research on Infectious Diseases of Poverty.*

Xiang, Z., Tian, Y., & He, Y. (2007). PHIDIAS: A Pathogen-Host Interaction Data Integration and Analysis System. *Genome Biology*(8), S. R150.

Zhang, X.-L., Jeza, V. T., & Pan, Q. (2008). *Salmonella Typhi*: from Human Pathogen to a Vaccine Vector. *Cellular & Molecular Immunology*(5:2), S. 91-97.