Informatik-gestützte Fallstudie zur Analyse von Proteinstrukturen am Beispiel der Pepsinresistenz von Interleukin-8

Dissertation

zur Erlangung des akademischen Grades

Dr. rer. med.

an der Medizinischen Fakultät

der Universität Leipzig

eingereicht von:

Gerd Anders

Geburtsdatum / Geburtsort:

15.10.1976 / Berlin

angefertigt am:

Institut für Medizinische Physik und Biophysik
der Universität Leipzig

Betreuer:

Prof. Dr. rer. nat. Daniel Huster

Beschluss über die Verleihung des Doktorgrads vom: 12.02.2019

Inhaltsverzeichnis

I	Einführung	6
1	Einleitung und Zielstellungen	7
1.1	Die Bedeutung der Strukturbio­logie in der Biomedizin	7
1.2	Proteinstrukturebenen	9
1.3	Zielstellungen dieser Disserta­tionsschrift	10
1.3.1	1) Aufbau einer lokalen In­frastruktur zwecks Manage­ment von Proteinstruk­turdaten	10
1.3.2	2) Entwicklung und Imple­mentation eines neuartigen Verfahrens zur Analy­se von Proteinstrukturen	10
1.3.3	3) Algorithmische Evalua­tion und visuelle Analyse von Proteinstrukturen	12
1.3.4	4) Experimentelle Verifi­kation der berechneten Ergebnisse	16
1.4	Zusammenfassung und Ausblick	16
	Anhänge	
Anhang 1.A	Das SCOP-Schema	19
Anhang 1.B	Weitere berücksichtigte Ver­fahren für SECSTREME	19
1.B.1	DSSP	19
1.B.2	Graphentheoretische Ver­fahren	19
1.B.3	Sonstige Verfahren	19
2	PDB2 - eine Infrastruktur für Daten der Protein Data Bank	20
2.1	Motivation	20
2.2	Die Protein Data Bank (PDB)	21
2.3	Die Datenbank-basierte PDB2-Infrastruktur	23
2.3.1	Die PDB Markup Language (PDBML)	23
2.3.2	Kurzeinführung in die Da­tenbank-basierte PDB2-Infrastruktur	23
2.3.3	Zusammenfassung und Ausblick	25
	Anhänge	
Anhang 2.A	Abbildungen	27
Anhang 2.B	PDB-bezogene Projektüber­sicht (Auswahl)	30
2.B.1	Projekte zur strukturellen Klassifizierung von Proteinen	30
2.B.2	Datenbanken bzw. Datenkollektionen von Proteinannotationen	30
2.B.3	Diverse Projekte des PDB-Konsortiums	30
Anhang 2.C	Nicht PDB-bezogene Pro­jekte (Auswahl)	31
2.C.1	Abfrage-/Speichersysteme	31
2.C.2	Softwareentwicklungen	31
2.C.3	Datenbank-theoretische Projekte	31

II	Ausgewählte Publikationen	32
III	Zusammenfassung	59
IV	Anhang	64
	Literaturverzeichnis	65
A	Erklärung über die eigenständige Abfassung der Arbeit	88
B	Beruflicher Werdegang	89
C	Publikationen	90
D	Danksagung	91

Tabellenverzeichnis

2.1	Anzahl verfügbarer PDB-Strukturen nach Verfahren (Stand: Ende Dezember 2017).	21
-----	---	----

Abbildungsverzeichnis

1.1	[Burley18]: Repräsentative PDB-Strukturen mit nachhaltigem Einfluss u. a. in der Biologie und Biomedizin. (a) Nucleosome Core Particle (PDB code: 1aoi; [Luger97]); (b) Major Histocompatibility Complex 1 (1hla; [Bjorkman87]); (c) Photosystem II (1s5l; [Ferreira04]).	8
1.2	Verschiedene Topologien (A, B) zw. α -Helices (H) und β -Strängen (S) ([Krissinel04]).	9
1.3	Schematische Illustration eines Vektors (grauer Pfeil), der eine Helix abstrakt repräsentiert ([Majumdar05]).	11
1.4	Schematische Struktur von IL8 (PDB Code: 1il8). Der N-/C-Terminus ist in blau/rot eingefärbt. Konservierte Cystine sind gelb markiert.	13
1.5	Schematische Struktur von PI3 (PDB Code: 1f34). Der N-/C-Terminus ist in blau/rot eingefärbt. Cystine sind gelb markiert. Für einige Residuen liegen keine 3D-Koordinaten vor (rechts oben, nahe der C-terminalen α -Helix).	14
2.1	Das historische PDB-Format am Beispiel des PDB-Datensatzes 1bbz (Auszug). . . .	20
2.2	Das Wachstum der PDB seit 1990 bzgl. hinterlegter Proteinstrukturen. Für jedes Halbjahr wird die Gesamtanzahl von allen (bzw. neu erfassten) Proteinstrukturen mit grauen (bzw. schwarzen) Balken dargestellt.	22
2.3	Das Wachstum der PDB seit 1990 bzgl. der Gesamtgröße aller PDBML-Dateien. Für jedes Halbjahr wird die Gesamtgröße aller (bzw. neu erfassten) PDBML-Dateien mit grauen (bzw. schwarzen) Balken dargestellt.	22
2.4	Das modernere PDBML-Format am Beispiel des PDB-Datensatzes 1BBZ (Auszug), s. auch Abb. 2.1.	24
2.5	Das DB2 XML-relationale Speichermodell (Quelle: [Bhambhri05]).	25
2.6	XML-Baum zu Abb. 2.4 auf Seite 24.	28
2.7	Unveröffentlichtes Schema einer PDB-Referenzdatenbank des PDB-Konsortiums (1990er Jahre).	29

Teil I

Einführung

Kapitel 1

Einleitung und Zielstellungen

Proteine sind Biomoleküle, die in allen Lebensformen, d. h. Organismen, vorkommen. Proteine sind Makromoleküle, die sich hinsichtlich ihrer Struktur stark unterscheiden. Aufgrund dieser Strukturvielfalt können Proteine sehr spezifische biochemische oder biophysische Funktionen ausführen. Ein wesentliches Dogma in der Molekularbiologie aus den 1970er Jahren besagt, dass die Funktion(en) eines Proteins sehr von der zugrunde liegenden Molekülstruktur abhängt ([Anfinsen77]). Die Kenntnis der Molekülstruktur kann folglich helfen, auf die Funktion(en) bzw. Wirkungsweise(n) eines Proteins zu schließen (s. u.). Die Strukturbiologie ist eine Disziplin der Grundlagenforschung, welche sich mit der Molekularstruktur von biologischen Makromolekülen befasst ([Banaszak00]). Das Annotieren der Struktur und Funktion(en) von Proteinen ist eine der wichtigsten Aufgaben in der Strukturbiologie und der Bioinformatik ([Tramontano05]).

1.1 Die Bedeutung der Strukturbiologie in der Biomedizin

Ist die Struktur bzw. die Funktionsweise eines Proteins bekannt, so können diese Informationen insbesondere in der Biomedizin hilfreich sein. Das wird stellvertretend an zwei Beispielen erläutert.

1958 wurde erstmalig die Atomstruktur eines Proteins - von Myoglobin - experimentell ermittelt ([Kendrew58], [Kendrew60]), später die Atomstruktur von Hämoglobin ([Perutz60], [Bolton70]). Hämoglobin ist ein eisenhaltiger Proteinkomplex, welcher aus vier Proteinen (Globinen) besteht. Es bindet und transportiert Sauerstoff in den Erythrozyten (rote Blutkörperchen). Durch die Kenntnis der Molekularstruktur konnte u. a. die reversible Funktionsweise (Be- und Entladen von Sauerstoff) von Hämoglobin aufgeklärt werden. Neben dem Verstehen dieses Zellprozesses kann aufgrund der Kenntnis der Proteinstruktur auch die Sichelzellanämie erklärt werden. Diese Krankheit wird durch eine Genmutation ausgelöst. Dabei wird eine strukturelle Eigenschaft des Hämoglobins signifikant verändert, welche sich negativ auf die Funktionalität des Proteins auswirkt.

Die Proteinstruktur vom Major Histocompatibility Complex 1 (MHC 1) wurde 1987 publiziert ([Bjorkman87]), s. Abb. 1.1. In der biomedizinischen Forschung hatte dies eine Signalwirkung: Wissenschaftler versuchten die Interaktionen von Antigenen und T-Zellen anhand der 3D Struktur zu verstehen. Retrospektiv bleibt festzuhalten, dass diese und Hunderte ähnliche Proteinstrukturen das Wissen um die biologische Wirkungsweise von T-Zellen grundlegend veränderten. Zuvor waren die Erkenntnisse der Funktion und Regulation von T-Zellen sehr unvollständig. Die Gesamtheit der o. g. Proteinstrukturen offenbarte nicht nur eindeutig die Mechanismen, wie das humane Immunsystem krankhafte Zellen erkennt, sondern trug auch entscheidend zur zielgerichteten Entwicklung von Strategien zur direkten Modulation der Immunantwort zu therapeutischen Zwecken bei ([Burley18]). Im Allgemeinen kann eine Proteinstruktur mit bekannten Atomkoordinaten eine Ausgangsbasis für die Entwicklung neuer Medikamente oder von neuen Ansätzen zur Behandlung von Krankheiten bilden.

Abbildung 1.1: [Burley18]: Repräsentative PDB-Strukturen mit nachhaltigem Einfluss u. a. in der Biologie und Biomedizin. (a) Nucleosome Core Particle (PDB code: 1aoi; [Luger97]); (b) Major Histocompatibility Complex 1 (1hla; [Bjorkman87]); (c) Photosystem II (1s5l; [Ferreira04]).

Proteine bestehen aus Hunderten oder Tausenden Atomen, so dass mit zunehmender Anzahl von Atomen die (3D) Struktur eines Proteinmoleküls sehr komplex ist. Diese Komplexität bzw. Strukturvielfalt wirkt sich auf das allgemeine Vorgehen bei der experimentellen Bestimmung von Proteinstrukturen aus. Initiativen wie die Protein Structure Initiative und die Structural Genomics Initiative hatten das Ziel, alle vorkommenden Proteinfaltungen experimentell zu ermitteln ([Skolnick00]) und in der Protein Data Bank (PDB, s. Kapitel 2) zu hinterlegen. Die PDB ist ein einzigartiges und weltweites Datenarchiv (Repository), welches umfangreiche 3D Strukturdaten von Proteinen zur freien Verfügung stellt ([Berman14]). Für die globale biologische und biomedizinische Forschung ist die PDB eine der beiden weltweit bedeutsamsten Repositorien und zählt zu den am häufigsten benutzten digitalen Datenquellen in der biomedizinischen Forschung ([Berman00b], [Burley17a]). Jene Initiativen erweiterten das Wissen um Proteinstrukturen ([Todd05], [Chandonia06], [Berman09], [Nair09], [Khafizov14]) und führten darüber hinaus zu neuen funktionellen Erkenntnissen auf biochemischer und molekularer Ebene. Dennoch ist die vorliegende riesige Menge an Proteinstrukturdaten noch nicht erschlossen, z. B. für die Biomedizin ([Hühne07], [Gales13]).

Mit der experimentellen Bestimmung einer Proteinstruktur werden die 3D-Atomkoordinaten eines Makromoleküls ermittelt. Die Gesamtheit aller bisher experimentell ermittelten 3D-Atomkoordinaten (Ende 2017: ca. 800 Millionen) stellt aus Sicht der (Bio-)Informatik insbesondere an die Speicherung (effizienter Zugriff, etc.) und algorithmische Datenverarbeitung (Programmlaufzeiten, usw.) von Proteinstrukturen nicht unerhebliche Anforderungen ([C97], [Gerstein99], [Albeck06]). Beispielsweise fokussierte sich das Hochleistungsrechnen (Supercomputing) zeitweise vor allem auf Anwendungen aus der Biologie ([Wadman99]), insbesondere aus der Strukturbiologie [Butler99], [Niiler01]), aber auch direkt auf Proteinstrukturen ([Lawrence01], [Sunyaev05]). Die gewonnenen Informatikerkenntnisse aus den (struktur-)biologischen Anwendungen wurden u. a. auf medizinische Anwendungen adaptiert und transferiert ([Bertone01], [Bayat02], [Bry03]), auch für die Behandlung von Krebs (s. u. a. [Demir17]).

Diese überwiegend numerische Sammlung von 3D-Atomkoordinaten ist eine wertvolle Datengrundlage für die weitere Struktur- und Funktionsanalyse des Proteins. Nur auf Basis dieser 3D-Daten ist jedoch die biologische Funktion nicht unmittelbar ableitbar, so dass ihr Nutzen zunächst eingeschränkt ist ([Gerstein00]). Deswegen werden weitere Verfahren angewendet mit dem Ziel, die Funktion(en) eines Proteins zu bestimmen. Ein speziell für den Vergleich von Proteinstrukturen entwickeltes Verfahren wird in Abschnitt 1.2 vorgestellt.

1.2 Proteinstrukturebenen

Vergleiche von Proteinstrukturen sind grundlegende Verfahren, welche u. a. in vielen Bereichen biomedizinischer Studien angewendet werden. Der Anwendungsbereich umfasst die Klassifizierung von Proteinen, die Vorhersage der Funktion(en) eines Proteins, das Aufdecken evolutionärer Beziehungen von Proteinen, usw. ([Abyzov07]). Strukturbiologen beschreiben oft besondere Merkmale einer Proteinfaltung durch die allgemeine Anordnung von α -Helices¹ und β -Faltblättern (sowie Schleifen), d. h. von Sekundärstrukturelementen.

Neben der Sekundär- und Tertiärstruktur gibt es noch weitere Proteinstrukturebenen, die wie folgt charakterisiert werden ([Lagerstrøm-Lang52]):

- Primärstruktur: Peptidbindungen verknüpfen Aminosäuren zu einer Proteinkette (Sequenz).
- Sekundärstruktur: Lokale, räumliche und sich wiederholende Elemente (α -Helices, β -Stränge) einer Sequenz aufgrund von Wasserstoffbrücken zwischen Residuen, d. h. zwischen verknüpften Aminosäuren.
- Tertiärstruktur: (kompakte) 3D-Struktur eines Proteins.
- Quartärstruktur: 3D-Struktur eines Protein(komplexe)s mit allen Sequenzen bzw. Untereinheiten (nicht anwendbar bei monomeren Proteinen).

Für Strukturbiologen repräsentieren Sekundärstrukturelemente in der Regel die Basiskomponenten für die Klassifizierung von Proteinstrukturen ([Lesk81]). Klassifizierungsschemata wie SCOP² ([Murzin95]) basieren auf der Topologie (auch als Konnektivität bezeichnet) von Sekundärstrukturelementen ([Dundas07]), s. Abb. 1.2 und Abschnitt 1.A im Anhang.

α -Helices und β -Stränge sind lokal, räumlich und regelmäßig wiederkehrende Sekundärstrukturelemente und kommen in fast allen bekannten Proteinstrukturen vor ([Murzin96], [Orengo99]). Sowohl α -Helices als auch β -Stränge werden jeweils durch Abschnitte von irregulären Strukturen voneinander getrennt. Diese Abschnitte werden als Schleife (loop) oder Windung (coil) bezeichnet und können von struktureller Bedeutung sein ([Scheeff03]). Schleifen repräsentieren eine (eigene) Sekundärstrukturkategorie ([Leszczynski86]), werden aber in dieser Arbeit nicht berücksichtigt. Aufgrund der besonderen (strukturellen) biologischen Bedeutung der α -Helices und β -Stränge als einzig reguläre Sekundärstrukturelemente werden beide auch als 'building blocks' bezeichnet. Diese building blocks sind eine zentrale Komponente dieser Dissertationsschrift, die im Folgenden kurz aufgegliedert wird.

Abbildung 1.2: Verschiedene Topologien (A, B) zw. α -Helices (H) und β -Strängen (S) ([Krissinel04]).

¹Neben α -Helices gibt es noch weitere Helixarten, die hier jedoch nicht berücksichtigt werden.

²Structural Classification of Proteins

1.3 Zielstellungen dieser Dissertationsschrift

Die Zielstellungen dieser Dissertationsschrift lassen sich in die folgenden vier Teilbereiche untergliedern.

1.3.1 1) Aufbau einer lokalen Infrastruktur zwecks Management von Proteinstrukturdaten

Wie in Kapitel 2 erläutert, wird grundsätzlich jede experimentell ermittelte Proteinstruktur in genau einer Datei gespeichert. Ende Dezember 2017 waren mehr als 136000 Dateien verfügbar. Würde man diese bspw. nach einfachen Stichwörtern wie der Art der verwendeten experimentellen Methode (s. Tabelle 2.1) durchsuchen, so wären Ende 2017 mehr als 31 Milliarden Zeilen einzulesen und zu vergleichen gewesen, obwohl insgesamt nur ca. 136000 Zeilen die relevanten Informationen enthalten. Das wird unten in den Abb. 2.1 und 2.4 illustriert. Das ist nicht nur hinsichtlich der experimentellen Methode sehr ineffizient, sondern gilt prinzipiell auch für beliebige andere Stichwortsuchen. Komplexere Abfragen, insbesondere auf 3D-Atomkoordinaten wie bspw. umfassende Abstands- und/oder Winkelberechnungen zwischen Atomen, lassen sich in der Praxis nicht bzw. kaum realisieren. Damit ist eine Datei-basierte 'Infrastruktur' für die Analyse von Proteinstrukturen prinzipiell völlig ungeeignet.

Im Kontext dieser Dissertationsschrift musste zuerst eine wesentlich geeignetere Infrastruktur in Betracht gezogen werden. In der Informatik befasst sich das Forschungsgebiet Datenbanken und Informationssysteme mit Aspekten der (effizienten) Speicherung und (schnellen) Abfrage riesiger Datenmengen. Die Wahl fiel auf eine noch zu entwickelnde und zu implementierende Datenbank-gestützte Infrastruktur, genannt PDB2, die vom Autor sehr individuell auf die PDB angepasst wurde. Mit SQL³ bzw. XQuery sind jedoch bereits komplexe Abfragen (s. o.) auf dem Gesamtdatenbestand der PDB2 möglich. Das eigentliche Potential - z. B. für strukturelle biologische Analysen - der Datenbank-basierten PDB2-Infrastruktur wird damit dennoch nicht annähernd ausgeschöpft. Ermöglicht wurde diese Abfragevariabilität durch eine damals noch weitestgehend unbekannte Form der Speicherung und Abfrage (s. Abschnitt 2.3) von Proteinstrukturen ([Anders11]).

1.3.2 2) Entwicklung und Implementation eines neuartigen Verfahrens zur Analyse von Proteinstrukturen

Um das Potential der PDB2-Infrastruktur strukturell biologisch zu erschliessen bzw. auszuschöpfen bedurfte es eines geeigneten Verfahrens, welches idealerweise an eine native Schnittstelle (API⁴) des Datenbanksystems angebunden wird. In diesem Zusammenhang dienten diverse und zahlreiche Applikationen als Inspiration für die Entwicklung einer eigenen strukturell biologischen Applikation. Als Ergebnis dieser Applikationsrecherchen wurde das algorithmisch anspruchsvolle geometrische Hashing (s. u.) als elementares Verfahren für SECSTREME⁵ ausgewählt. Eine vollständige Übersicht zu weiteren verwendeten Verfahren von SECSTREME befindet sich im Anhang, s. Abschnitt 1.B.

Das geometrische Hashing operiert auf den 3D-Atomkoordinaten der Sekundärstrukturelemente, d. h. auf α -Helices und β -Strängen. Aufgrund ihrer Formgebung sind α -Helices und β -Stränge prädestiniert für Proteinstrukturvergleiche. Allerdings ist die Berechnung von Proteinstrukturüberlagerungen (engl. alignment) aus algorithmischer Sicht rechenintensiv, was sich insbesondere bei Ausführungen gegen eine Vielzahl von Proteinstrukturen (große Strukturbibliotheken, die PDB, etc.) deutlich bemerkbar macht. Insbesondere erreichen Strukturüberlagerungen um z. B. eine einzelne

³Structured Query Language

⁴Application Programming Interface

⁵Die Bezeichnung SECSTREME (SECondary STRucture Elements MinEr) stammt vom Verfasser dieser Dissertation.

Proteinstruktur mit der gesamten PDB zu vergleichen keine befriedigende (Gesamt-)Laufzeit oder benötigten bspw. reichlich Rechenressourcen ([Ferrari03]).

1.3.2.1 Geometrisches Hashing, Abgrenzung zu anderen Verfahren

Aus diesem Grund wurden zahlreiche verschiedene Verfahren ([Carugo06], [Aung07]) entwickelt, um Überlagerungen von Proteinstrukturen 'sofort (instantly)' zu berechnen. Einige Verfahren wie bspw. VAST ([Madej95], [Gibrat96]) bilden dazu Sekundärstrukturelemente als Vektoren ab (s. Abb. 1.3). Einerseits führt dieser Ansatz zu einer teilweise erhebliche Datenreduktion (nur ein 3D-Koordinatenpaar pro Vektor). Andererseits ermöglicht diese Datenreduktion eine Datenaufbereitung in Form von Vorberechnungen, um die eigentliche Berechnung der Alignierung weiter zu verkürzen, z. B. durch (geometrische) Filterkriterien wie der Länge eines Vektors, etc.

Für die Entwicklung von SECSTREME waren aber nicht ausschliesslich Laufzeitaspekte maßgeblich, sondern die Frage, ob man geometrisches Hashing grundsätzlich auf sämtliche Proteinstrukturen in der PDB2-Datenbank anwenden kann. Geometrisches Hashing ([Wolfson97]) ist ein hoch-effizientes Verfahren zum Auffinden von geometrischen Merkmalen in einer Datenbank mit ebendiesen. Die besondere Stärke von diesem Verfahren in der Strukturbiologie ist das Finden kleiner, aber hochgradig ähnlicher 3D Substrukturen in Proteinen ([Pennec94], [Wallace97], [Pennec98]). Geometrisches Hashing wurde zwar bereits auf Sekundärstrukturelemente angewendet, jedoch nur auf eine kleine Datenmenge ([Holm95]).

Abbildung 1.3: Schematische Illustration eines Vektors (grauer Pfeil), der eine Helix abstrakt repräsentiert ([Majumdar05]).

Um das volle Potential des PDB2-Datenbestandes nutzbar zu machen, wurde nicht nur auf eine übliche Datenreduktion (z. B. Darstellung eines beliebig langen Sekundärstrukturelements als Vektor mit genau einem 3D Koordinatenpaar) bewusst verzichtet. Im Gegenteil, der Datenbestand wurde noch erhöht, um strukturbiologisch präzisere Abfragen mit praktisch maximaler Genauigkeit zu ermöglichen. Nur dadurch konnte (struktur-)biologisch eine seit Jahren offene Frage zwei Proteine (s. u.) betreffend umfassend auf molekularer Ebene beantwortet werden (s. [Anders16] in Teil II auf Seite 50).

Geometrisches Hashing als (Laufzeit-)kritischer Algorithmus musste aufwendig auf die berechneten Sekundärstrukturelemente angepasst und implementiert werden. Hintergrund dieses Vorgehens war die in der wissenschaftlichen Literatur vorherrschende Meinung, dass ein solcher algorithmischer Ansatz bereits auf größere Datenbestände nicht anwendbar ist ([Kiuchi07]). Folglich stellte sich im Kontext dieser Dissertationsschrift erst einmal die wissenschaftliche Frage, ob SECSTREME die angegebenen Beschränkungen abmildern oder sogar aufheben kann. Diese Frage sollte mit Evaluierungen von ausgewählten Proteinstrukturen beantwortet werden (s. Abschnitt 1.3.3). Die zahlreichen Ergebnisse, die SECSTREME als strukturbiologische Daten (Überlagerungen von Proteinstrukturen im PDB Format) in Dateiform ausgibt, lassen sich visuell analysieren (s. bspw. Abb. 2 in [Anders11]). Als Ergebnis dieser Evaluation konnte eindeutig nachgewiesen werden, dass SECSTREME (struktur-)biologisch relevante Daten generiert.

Seit 2016 ist eine Patentanmeldung für SECSTREME anhängig.

1.3.3 3) Algorithmische Evaluation und visuelle Analyse von Proteinstrukturen

Zum Zeitpunkt der Evaluierung von SECSTREME lag ein besonderes Forschungsinteresse auf der Proteinstruktur von Interleukin-8 (IL8) ([Pisabarro06]), welches zur Gruppe der Chemokine (*chemotaktische Zytokine*) gehört. Aus diesem Grund wurden vier Sekundärstrukturelemente (drei β -Stränge und die C-terminale α -Helix) als Eingabe für die Evaluierung verwendet. PI3 war eine von tausenden Proteinstrukturen⁶, die bzgl. einer Datenbanksuche im Evaluationskontext von SECSTREME gefunden wurden. Durch die visuelle Analyse der überlagerten Proteinstrukturen von IL8 und eines Pepsin-PI3-Proteinkomplexes konnten zahlreiche und für die Funktion von PI3 kritische Residuen bzw. Regionen auch in IL8 identifiziert werden. Dadurch konnten nicht nur die strukturellen Schlüsselstellen von IL8 aufgedeckt, sondern auch eine plausible Hypothese für die Pepsinresistenz von IL8 aufgestellt werden.

1.3.3.1 Chemokine

Chemokine sind eine große Unterfamilie der Zytokine und die größte Familie von Zytokinen in der humanen Immunophysiologie ([Rajaratnam01], [Bonecchi09]). Chemokine sind auch eine Familie von funktionell verwandten, kleinen, löslichen, globulären und sekretierten Proteinen, welche G-Protein-gekoppelte Rezeptoren aktivieren. In 2014 waren 22 Rezeptoren und ungefähr 45 Chemokine bekannt, die human exprimiert werden ([Bachelier14], [Kufareva17]), in 2015 bereits 23 bzw. ca. 50 ([Alexander15], [Berkamp17]). Darüber hinaus nehmen sie eine Schlüsselrolle u. a. in der Homöostase, Pathogenese von Tumoren und Automimmunkrankheiten ein ([Raman11]). Auch aus therapeutischer Sicht sind Chemokine interessante Zielmoleküle ([Garin11]).

Obwohl die meisten Chemokine gemeinsame funktionelle Eigenschaften haben, sind für ihre Familienzugehörigkeit einzig strukturelle Kriterien ausschlaggebend. Chemokine bestehen aus einzelnen Polypeptidketten mit jeweils 70-100 Residuen. Diese Sequenzen weisen untereinander eine Identität von 20%-95% auf und besitzen vier charakteristische Cysteine. Diese Cysteine sind konserviert und invariant, aber vor allem essentiell für die Struktur und Funktion (s. u.). Aufgrund der charakteristischen Eigenschaften (vor allem Sequenzpositionen) dieser vier Cysteine wurde offiziell eine Nomenklatur eingeführt, welche Chemokine in vier Unterfamilien einteilt ([Murphy00], [Zlotnik00], [Bacon02]). Basierend auf dem Sequenzabschnitt im Bereich der beiden N-terminalen Cysteine ([Fernandez02], [von Adrian03]) unterteilen sich diese vier Unterfamilien in je zwei Haupt- und Nebenfamilien ([Kelner94], [Bazan97]).

Hauptfamilien:

- CXC-Chemokine (α -Chemokine):
Bei dieser großen Chemokinfamilie sind die beiden N-terminalen Cysteine durch eine dazwischenliegende Residue voneinander separiert.
- CC-Chemokine (β -Chemokine):
Bei dieser großen Chemokinfamilie liegen die beiden N-terminalen Cysteine nebeneinander.

Nebenfamilien:

- (X)C-Chemokine (γ -Chemokine, Lymphotaktin)
- CX₃C-Chemokine (δ -Chemokine, Fraktalkine, Neurotaktin)

⁶Die vollständige Ergebnisliste ist der Originalpublikation ([Anders11]) beigelegt und digital verfügbar

Beide Unterfamilien weisen chemotaktische Besonderheiten auf ([Kologlu01]). Historisch bedingt, werden Chemokine auch in andere Unterfamilien eingeteilt, z. B. in Interleukine (IL-*n*). Eines dieser Interleukine ist IL8 und in Bezug auf die Nomenklatur gehört IL8 der Unterfamilie der CXC-Chemokine an.

1.3.3.2 Interleukin-8 (IL8)

IL8 (s. Abb. 1.4) wurde 1987 als ein neuartiger Typ eines neutrophil-aktivierenden Zytokins entdeckt ([Baggiolini89]). Eine der bemerkenswertesten Eigenschaften von IL8 ist die Expressionsbandbreite. In gesunden Geweben ist IL8 kaum vorhanden, jedoch kann die Menge rasant auf das zehnfache ansteigen ([Hoffmann02]).

Abbildung 1.4: Schematische Struktur von IL8 (PDB Code: 1il8). Der N-/C-Terminus ist in blau/rot eingefärbt. Konservierte Cystine sind gelb markiert.

IL8 ist ein sekretiertes Protein dessen Signalpeptid aus zwanzig Residuen besteht. Ohne dieses Signalpeptid verkürzt sich die Polypeptidkette von IL8 auf ca. 60 bis 95 Residuen. Die Sequenzidentität aller bekannten IL8-Proteine (ohne Signalpeptid) variiert von weniger als 20% bis über 90% ([Baggiolini92], [Fernandez02]). In der vorliegenden Dissertationschrift wird das Signalpeptid nicht berücksichtigt.

IL8 besteht aus fünf Sekundärstrukturelementen. Diese umfassen einen aus drei β -Strängen bestehendes anti-paralleles β -Faltblatt. Diese Anordnung ist ein bekanntes Super-Sekundärstrukturelement mit Namen Greek key ([Hutchinson93]). Das vierte Sekundärstrukturelement ist eine lange C-terminale α -Helix. Der N-Terminus enthält keine regulären Sekundärstrukturelemente mit Ausnahme einer kurzen 3_{10} -Helix, dem fünften und letzten Sekundärstrukturelement. Dieser Helix folgt direkt der erste β -Strang ([Clare95]). Nicht nur die strukturelle Rolle der Sekundärstrukturelemente hinsichtlich ihrer Funktion wurde auf molekularer Ebene detailliert studiert und analysiert (u. a. [Clark-Lewis94], [Clark-Lewis95], [Baggiolini97]). Darüber hinaus auch Unterschiede zwischen der Struktur in gelösten und kristallinen Zustand ([Baldwin91]).

Sekretierte globulare Proteine wie IL8 enthalten oft durch Disulfidbrücken verbundene Cysteinepaare (Cystine). Da Disulfidbrücken essentiell für die Faltung, Stabilität und Tertiärstruktur sind, sind Cystine in der Sequenz und im Raum konserviert. Aufgrund der strukturellen Bedeutung von

Abbildung 1.5: Schematische Struktur von PI3 (PDB Code: 1f34). Der N-/C-Terminus ist in blau/rot eingefärbt. Cystine sind gelb markiert. Für einige Residuen liegen keine 3D-Koordinaten vor (rechts oben, nahe der C-terminalen α -Helix).

Disulfidbrücken für IL8 wurden diese intensiv studiert und analysiert (u. a. [Thornton81], [Nagano99], [Bhattacharyya04]), ebenso wie die konservierten Cysteine ([Thangudu08]).

Beide Disulfidbrücken in IL8 übernehmen eine essentielle strukturelle und funktionelle Aufgabe ([Rajaratnam99]): Ein Entfernen von einer der beiden Disulfidbrücken resultiert u. a. in der Denaturierung von IL8. D. h., die Disulfidbrücken sind für die biologische Aktivität von IL8 essentiell ([Baggiolini94]).

1.3.3.3 Aspartylproteasen

Eines der interessantesten Ergebnisse von SECSTREME war eine Überlagerung von IL8 mit einem Aspartylproteaseinhibitor, PI3, s. Abb. 1.5. Aspartylproteasen ([Szecsi92]) sind Enzyme, die Peptidbindungen ihrer Substrate spalten können. Der Name ist von den (üblicherweise zwei) hochkonservierten Asparaginsäuren (auch Aspartate genannt) im aktiven Zentrum der Protease abgeleitet. Aspartylproteasen verwenden ein Wassermolekül, welches zusammen mit den Aspartaten die Katalyse der Peptidsubstrate bewirkt. Ihre optimale Funktion erreichen Aspartylproteasen in einer sauren Umgebung (z. B. im Magen eines Säugetiers). Analog zu den Disulfidbrücken von IL8, sind die Disulfidbrücken, welche von konservierten Cysteinen gebildet werden, ebenfalls charakteristische Merkmale von Aspartylproteasen. Vertreter dieser katalytischen Enzyme sind u. a. Pepsin und Cathepsin (s. u.).

1.3.3.4 Pepsin Inhibitor-3 (PI3)

Proteine, die Aspartylproteasen inhibieren, sind in der Natur sehr selten ([Girdwood00]). Der Fadenwurm *Ascaris suum* ist ein sehr häufig vorkommender Darmparasit von (Haus-)Schweinen. *A. suum* ist dabei Verdauungsenzymen wie den Proteasen seines Wirts ausgesetzt. Zu seinem Schutz produziert der Fadenwurm einige Proteaseinhibitoren ([Ng00]). Einige dieser Inhibitoren von *A. suum* sind als Pepsin-spezifisch bekannt ([Davies90], [Fruton02], [Dunn02]). Der am häufigsten vorkommende Inhibitor ([Abu-Erreish74]) ist ein Aspartylproteaseinhibitor, PI3. PI3 wurde sequenziert ([Martzen90]), geklont ([Kageyama98]) und ist einer der wenigen Inhibitoren von Aspartylproteasen mit bekannter Molekularstruktur ([Li00]). PI3 inhibiert reversibel sowohl die proteolytischen Aktivitäten der Magenenzyme Pepsin und Gastricin ([Abu-Erreish74]), als auch die von Cathepsin E ([Keilova72], [Samloff87], [Jupp88]) jeweils reversibel. Andere Aspartylproteasen ([Valler85], [Kageyama98]) wie humanes Trypsin ([Hawley94], [Hawley95]) werden nicht signifikant inhibiert.

Proteaseinhibitoren sind Single- oder Multi-Domänenproteine (Interpro ([Hunter09]) IPR010480)). PFAM ([Finn10]) verzeichnet für PI3 zwei Proteindomänen. Aufgrund der Sequenzwiederholung handelt es sich dabei um ein Tandem, d. h. um eine Domänenwiederholung (repeated domain) (PFAM ID: PF06394).

Unabhängig davon hat PI3 keine signifikante Sequenzhomologie mit anderen bekannten Proteinen, außer Proteinen von Fadenwürmern⁷ ([Willenbücher93], [Ng00]). Die Sequenz von PI3 besteht aus 149 Residuen und enthält sechs Cysteine. Davon sind vier konservierte Cysteine, die Disulfidbrücken ausbilden ([Martzen90], [Zalatoris99]). Die verbliebenen beiden Cysteine bilden eine nicht-konservierte, aber wahrscheinlich notwendige Disulfidbrücke aus ([Girdwood00], [Ng00]).

PI3 hat eine sehr ungewöhnliche Faltung von zwei anti-parallelen β -Faltblättern. Jedes Faltblatt wird von einer α -Helix flankiert, sodass sich von der Anordnung der Sekundärstrukturelemente auch die oben erwähnte Domänenaufteilung ergibt ([Ng00]).

1.3.3.5 Funktionelle Wechselwirkungen zwischen Pepsin und Interleukin-8

Aus immunologischer bzw. biologischer Sicht mag es nicht offensichtlich sein, dass die Pepsinresistenz von IL8 durchaus von funktioneller Bedeutung ist. Neugeborene konsumieren beim Stillen über die Muttermilch IL8, welches in den Magen des Säuglings gelangt ([Polat16]). Dort würde es ohne 'intrinsic pepsin resistance' ([Maheshwari05]) von Pepsin gespalten und verdaut werden, sodass es beim Aufbau des Immunsystems nicht helfen könnte. Für die strukturellen Komponenten von PI3, welche für die Pepsininhibition maßgeblich sind, konnten mit Hilfe von SECSTREME weitestgehend strukturelle Entsprechungen in IL8 identifiziert werden. Die Gesamtheit dieser Entsprechungen ergab erstmals eine plausible Hypothese für die Pepsinresistenz von IL8 auf molekularer Ebene. Diese Hypothese lieferte wertvolle Anhaltspunkte für die Verifizierung der Hypothese durch verschiedene Laborversuche (s. Teil II), wobei die Ergebnisse der Experimente in allen wesentlichen Punkten im Einklang mit der Hypothese standen.

In der wissenschaftlichen Literatur gibt es nennenswerte Beobachtungen bzw. Erkenntnisse zu Pepsin, IL8 bzw. PI3 im immunologischen Kontext. Allerdings konnten diese meist vagen Informationen für die strukturelle Analyse dieser Proteinstrukturen nicht verwertet werden. Dennoch ergaben diese Informationen ein stimmiges Bild hinsichtlich der Hypothese, sodass im Folgenden ein grober Überblick gegeben wird.

PI3 verursacht als Pepsininhibitor Immunreaktionen ([Kageyama98]), welche auf eine immunomodulatorische Funktion hindeuten ([Krushna11]). In Anbetracht der zentralen Rollen von Proteasen hinsichtlich der angeborenen und erworbenen Immunreaktionen ist es überraschend, dass solchen Inhibitoren nicht mehr Aufmerksamkeit geschenkt wurde ([Knox07]): Nachweise, dass Proteaseinhibitoren von Parasiten Immunreaktionen herunterregeln können, gab es bereits Mitte der

⁷Von all diesen Proteinen ist lediglich PI3 als Proteaseinhibitor bekannt ([Ng00]).

1980er Jahre ([Ribeiro85], [Wikel86], [Leid87a], [Leid87b]). Einige Studien zeigten, dass nicht nur wenige Fadenwürmer, sondern auch Endosymbionten von Fadenwürmern ([Bazzocchi03]) Immunreaktionen - mit Produktion von IL8 - verursachen ([Bazzocchi03], [Gause03], [Proudfoot04]). Andere Studien offenbarten bemerkenswerte immunologische Schutzmechanismen ([Finkelman97]). Alle diese Studien bieten jedoch keine strukturellen Anhaltspunkte, warum IL8 hochgradig resistent gegenüber (humanem) Pepsin ist und in Assays seine strukturelle, immunologische und funktionelle Integrität beibehält ([Maheshwari05]).

Die Überlagerung der Sekundärstrukturelemente von IL8 und PI3 mit ihren unterschiedlichen Topologien mag auch evolutionär von Interesse sein ([Abagyan89], [Efimov95], [Heinemann95], [Alexandrov96a], [Alexandrov96b], [Gibrat96], [Lindqvist97], [Przytycka99], [Fliess02], [Yuan05]) z. B. für die Entstehung der Tandemdomänen oder die geringe Sequenzidentität von PI3 (s. o.). Allerdings waren evolutionäre Aspekte nicht Untersuchungsgegenstand dieser Dissertation.

1.3.4 4) Experimentelle Verifikation der berechneten Ergebnisse

Gegenwärtig stützen sich Strukturbologen vermehrt auf ergänzende experimentelle Messungen um Forschungsergebnisse zu verbessern ([Burley17b]). Durch eine detaillierte visuelle Analyse der überlagerten Proteinstrukturen von IL8 und PI3 ergaben sich sehr konkrete Anhaltspunkte, um zielgerichtet geeignete Laborexperimente zwecks Validierung der Hypothese auszuwählen.

Nach der visuellen Analyse (s. Abschnitt 1.3.3) wurden mehrere Laborexperimente zwecks Überprüfung der Hypothese durchgeführt, wobei der Verfasser dieser Dissertationsschrift an sämtlichen Laborversuchen nicht aktiv beteiligt war.

1.4 Zusammenfassung und Ausblick

Die primäre Aufgabenstellung war die Konzeption und Umsetzung einer geeigneten Infrastruktur mit dem Ziel, den Gesamtdatenbestand der PDB für strukturelle Analysen zu erschliessen. Als Ergebnis wurde eine speziell auf die PDB angepasste Datenbank-gestützte PDB2-Infrastruktur aufgebaut.

Die sekundäre Aufgabenstellung war die Entwicklung eines Prototyps von SECSTREME, um insbesondere die Anwendbarkeit von geometrischem Hashing zu evaluieren. Als Eingabe wurden vier Sekundärstrukturelemente von IL8 verwendet. Die damit verbundene Datenbanksuche ergab tausende Funde von Proteinstrukturen mit ähnlicher räumlicher Anordnung von Sekundärstrukturelementen. Stichprobenartige visuelle Überprüfungen der Ergebnisse der Datenbanksuche ergaben ausnahmslos sinnvolle strukturblogische Überlagerungen der Sekundärstrukturelemente. Aufgrund dieser strukturblogisch vielversprechenden Überlagerungen ist die Verwendung von geometrischem Hashing für die Suche nach ähnlichen Anordnungen von Sekundärstrukturelementen pragmatisch.

Die o. g. stichprobenartige Evaluierung von Überlagerungen umfasste u. a. auch einen Proteinkomplex bestehend aus Pepsin und einem Pepsininhibitor-3. Diese Überlagerung wurde strukturblogisch analysiert. Die Analyse resultierte in einer Hypothese, welche die ungewöhnliche Pepsinresistenz von IL8 plausibel auf molekularer Ebene erklären konnte.

Um diese Hypothese zu verifizieren, wurden diverse Laborexperimente durchgeführt. Es konnte nachgewiesen werden, dass IL8 von Pepsin nicht gespalten wird, jedoch inhibiert es Pepsin nicht. Denaturiertes IL8 wurde in Versuchen gespalten, sodass die Struktur von IL8 für die Pepsinresistenz bedeutsam ist. In weiteren Experimenten wurden die Wechselwirkungen von Pepsin und IL8 untersucht. Damit konnte die Region von IL8 eingegrenzt werden, welche mit Pepsin interagiert.

Neben PI3 gab es nach Auffassung des Autors noch weitere strukturblogisch interessante Ergebnisse in Form von Strukturüberlagerungen mit IL8. Diese überlagerten Proteinstrukturen jeweils

analog zu PI3 einer Detailanalyse zu unterziehen würde den Rahmen dieser Dissertation bei weitem sprengen.

PI3 interagiert mit Pepsin. IL8 (als Ligand) interagiert sowohl mit seinem Rezeptor als auch mit Pepsin. Die Frage, ob folglich auch Pepsin mit einem Interleukinrezeptor agiert, wurde an überlagerten Proteinstrukturen untersucht. Diese Überlagerungen lassen strukturellbiologisch nicht darauf schließen, dass eine derartiges Interagieren plausibel ist.

Kapitelanhang

1.A Das SCOP-Schema

Um Anhaltspunkte für die Funktionsweise von Proteinen abzuleiten, ist es nicht nur notwendig Proteine in biologisch sinnvolle Kategorien zu klassifizieren, sondern auch mit anderem biologischen Wissen anzureichen ([Gerstein00], [Ouzounis03]).

Die PDB verwaltet grundsätzlich Primärdaten zu Proteinstrukturen ([Berman08]). Eine geeignete strukturelle Klassifikation der bereits bekannten Proteine ist für deren erfolgreiche Analyse notwendig ([Gerstein00]). Ein weit verbreitetes Schema zur Klassifizierung von PDB-Proteinstrukturen ist SCOP ([Murzin95], [Hubbard97]). SCOP (Structural Classification of Proteins) klassifiziert Proteinstrukturen nach vier hierarchischen Ebenen. Diese diskreten Ebenen sind *class*, *common fold*, *superfamily* und *family*. Die Faltung (*common fold*) kann dabei als Anordnung der Sekundärstrukturelemente im 3D-Raum definiert werden, unter Berücksichtigung ihrer topologischen Konnektivität.

1.B Weitere berücksichtigte Verfahren für SECSTREME

1.B.1 DSSP

SECSTREME verwendet ein Programm mit Namen DSSP (Dictionary of protein secondary structure) zur Bestimmung der Sekundärstrukturelemente (α -Helices und β -Stränge). DSSP ([Kabsch83], [Touw15]) wird de facto als Verfahren der Wahl zur Bestimmung von Sekundärstrukturelementen angesehen ([Martin05]), trotz möglicher Schwierigkeiten bei Proteinstrukturen mit schwacher Auflösung ([Labesse97], [Martin05]).

1.B.2 Graphentheoretische Verfahren

Graphentheoretische Verfahren wie Clique Detection ([Butenko06]) werden oft für Vergleiche von Proteinstrukturen verwendet ([Mitchell89], [Grindley93], [Mizuguchi95], [Koch96], [Harrison02], [Weskamp04], [Kolbeck06], [Gardiner97]). Proteinstrukturen werden häufig als Vektoren dargestellt ([Abagyan88], [Abagyan89]). Ein Vektor symbolisiert dabei ein Sekundärstrukturelement, was zu einer vereinfachten (Datenreduktion) und vereinheitlichten Darstellung führt. Trotz Datenreduktion ist die Menge der Eingabedaten für graphentheoretische Verfahren wie Clique Detection kritisch in Bezug auf die Laufzeitkomplexität: Clique Detection ist ineffizient, da NP-vollständig ([Karp72]). D. h., erhöht sich die Menge der Eingabedaten (hier Sekundärstrukturelemente), erhöht sich die Gesamtlaufzeit überproportional, also drastisch. Als Konsequenz werden Verfahren wie Clique Detection eher auf kleinen Molekülen bzw. Datenmengen angewendet (z. B. [Willett95]).

SECSTREME verwendet für Clique Detection die Anwendung Cliquer ([Niskanen03]).

1.B.3 Sonstige Verfahren

Ergab eine Datenbanksuche passende Sekundärstrukturelemente, so sind diese für eine weitergehende Analyse noch nicht optimal aufbereitet. Eine entsprechende Aufbereitung wäre eine Überlagerung mit den Sekundärstrukturelementen aus der Eingabe, sodass diese Überlagerung visuell analysiert werden kann.

Es gibt verschiedene Verfahren für die optimale Überlagerung von zwei Punktmengen (hier Atomkoordinaten der Sekundärstrukturelemente) im 3D-Raum ([Carugo07]). SECSTREME verwendet dafür LAPACK ([Anderson99]). LAPACK (Linear Algebra PACKage) ist eine mathematische Bibliothek. SECSTREME verwendet von LAPACK eine Matrixfaktorisierung mit Namen Singular Value Decomposition (s. [Stewart93] für einen Review) und ersetzt dadurch ein zuvor verwendetes Verfahren ([Kabsch76]).

Kapitel 2

PDB2 - eine Infrastruktur für Daten der Protein Data Bank

2.1 Motivation

Digitale Daten sind leicht kopierbar, schnell übertragbar, etc. - und eine wirkliche Herausforderung ([Davidson95]). Allerdings existieren viele verschiedene - und teilweise hoch-komplexe - Datenformate, die nicht selten inkompatibel zueinander sind. Daher müssen Bioinformatiker viel Zeit mit der Konvertierung und Integration von Daten aus solchen Formaten aufwenden ([Köhler04]). Ein solches Format ist das ursprüngliche - d. h. historische - Datenformat der Protein Data Bank (PDB). Die PDB wurde im Jahr 1971 als Datei-basiertes Archiv für makromolekulare Strukturen gegründet ([Bernstein77]). Seit mehr als 40 Jahren ([Berman12], [Burley13]) ist die PDB ein globales Datenarchiv von weltweiter Bedeutung ([Burley18]), an das bspw. Laboratorien weltweit ihre experimentell ermittelten 3D-Atomkoordinaten von Proteinen u. a. Makromolekülen übermitteln.

Die Gebrauchstauglichkeit des ursprünglichen PDB-Formats ist eingeschränkt, insbesondere für PDB Datenkonsumenten. Intensiver Konsument ist fast jedes pharmazeutische and biotechnologische Unternehmen, welches in der Arzneimittelforschung tätig ist ([Burley17b]). Das PDB-Format ist insbesondere historisch bedingt ([Callaway96], [Berman08]). Dieses Format wurde anhand der Eigenschaften von Lochkarten mit 80 Spalten entwickelt. Eine solche Lochkarte war damals für das Speichern aller Strukturdaten eines PDB-Datensatzes - wie bspw. 1bbz - in einer einzelnen Datei notwendig. Das PDB-Format ist auszugsweise (sogenannter Header) in Abb. 2.1 illustriert.

Weil eine Vielzahl an Anwendungen von diesem Format abhängen, hat es sich in den letzten

```
HEADER  COMPLEX (TRANSFERASE/PEPTIDE) 28-APR-98 1BBZ
TITLE   CRYSTAL STRUCTURE OF THE ABL-SH3 DOMAIN COMPLEXED WITH A
TITLE   2 DESIGNED HIGH-AFFINITY PEPTIDE LIGAND: IMPLICATIONS FOR
TITLE   3 SH3-LIGAND INTERACTIONS
KEYWDS  COMPLEX (TRANSFERASE/PEPTIDE), SIGNAL TRANSDUCTION, SH3
KEYWDS  2 DOMAIN
EXPDTA  X-RAY DIFFRACTION
AUTHOR  M.T.PISABARRO,L.SERRANO,M.WILMANN
```

Abbildung 2.1: Das historische PDB-Format am Beispiel des PDB-Datensatzes 1bbz (Auszug).

Jahrzehnten nur unwesentlich verändert. Für einen zeitgemäßen Gebrauch - z. B. im Hinblick auf Data Mining - ist dieses ursprüngliche PDB-Format mit seinen semi-strukturierten Daten und seiner beachtlichen Komplexität nur schwer zu verarbeiten, d. h. zu parsen. Aus diesem Grund wird die PDB gelegentlich auch als "worst case scenario of biological data representation" ([Rother04]) bezeichnet.

Das Einführen eines alternativen oder erweiterten PDB-Formats als (Übergangs-)Lösung birgt gewisse Risiken: Die Anforderung nach einem neuen, moderneren Datenformat muss einerseits gegen den Zusatzaufwand (Formatpflege, Inkonsistenzprüfungen, usw.) und die historisch gewachsene Datenstabilität des gesamten Datenarchivs andererseits abgewogen werden. Gleichwohl arbeitet das PDB-Konsortium unentwegt an Verbesserungen bezüglich der Datenqualität bzw. -integrität (Bereinigung von Inkonsistenzen) der archivierten Proteinstrukturen ([Bhat01], [Westbrook02]), inklusive kompletter Überarbeitungen ([Dutta05], [Henrick08], [Lawson08]). In diesem Kontext wurde auch ein neues Datenformat mit Namen PDBML (Protein Data Bank Markup Language) eingeführt, welches in Abschnitt 2.3.1 kurz vorgestellt wird.

Dieses neue Datenformat ist für die vorliegende Dissertationsschrift von entscheidender Bedeutung. Es ermöglicht die vollständige Integration sämtlicher PDB-Daten in eine Datenbank (hier DB2 von IBM für Linux, Unix und Windows). Im Folgenden werden zunächst die PDB und das neue Datenformat vorgestellt. Anschliessend wird die Datenbank-gestützte PDB2-Infrastruktur kurz erläutert, welche für die spätere Analyse von Proteinstrukturen eine zentrale Komponente ist.

2.2 Die Protein Data Bank (PDB)

Vor rund 25 Jahren bestand das PDB-Archiv aus vergleichsweise wenigen, d. h. lediglich Hunderten von Proteinstrukturen. Seit der "Datenexplosion 1989-1992" ([Sussmann99], [Berman08]) ist die PDB jedoch beständig und rasant in der Anzahl der hinterlegten und verfügbaren Proteinstrukturen gewachsen. Im Jahr 2011 waren es mehr als 100000 Proteinstrukturen und wöchentlich/halbjährlich kommen Dutzende/Tausende neue hinzu. Dieses exponentielle Wachstum ist sowohl in Abb. 2.2 illustriert, als auch in Abb. 2.3 hinsichtlich der Gesamtgröße (in Gigabytes). So wurden in den ersten sechs Monaten 2014 mehr als 4600 Proteinstrukturen hinzugefügt (schwarzer Balken), sodass insgesamt mehr als 100000 hinterlegte Proteinstrukturen (grauer Balken) am Ende dieser Periode vorhanden waren (Abb. 2.2). Die Ende Dezember 2017 rund 136000 verfügbaren Proteinstrukturen umfassen insgesamt ca. 1430 Gigabyte an Daten (Abb. 2.3), wobei die 1 TB-Marke 2014 überschritten wurde.

Tabelle 2.1 listet sämtliche experimentellen Verfahren auf, mit denen diese 136584 Proteinstrukturen ermittelt wurden. Ende Dezember 2017 wurden mehr als 98% aller Proteinstrukturen durch zwei Verfahren bestimmt: Überwiegend durch die Röntgenkristallographie (X-ray diffraction;

Tabelle 2.1: Anzahl verfügbarer PDB-Strukturen nach Verfahren (Stand: Ende Dezember 2017).

(experimentelles) Verfahren	Anzahl	(experimentelles) Verfahren	Anzahl
FLUORESCENCE TRANSFER	1	ELECTRON CRYSTALLOGRAPHY	71
INFRARED SPECTROSCOPY	4	SOLID-STATE NMR	106
THEORETICAL MODEL	7	NEUTRON DIFFRACTION	125
EPR	8	ELECTRON MICROSCOPY	1890
POWDER DIFFRACTION	19	SOLUTION NMR	12016
FIBER DIFFRACTION	39	X-RAY DIFFRACTION	122248
SOLUTION SCATTERING	50	Σ	136584

Abbildung 2.2: Das Wachstum der PDB seit 1990 bzgl. hinterlegter Proteinstrukturen. Für jedes Halbjahr wird die Gesamtanzahl von allen (bzw. neu erfassten) Proteinstrukturen mit grauen (bzw. schwarzen) Balken dargestellt.

Abbildung 2.3: Das Wachstum der PDB seit 1990 bzgl. der Gesamtgröße aller PDBML-Dateien. Für jedes Halbjahr wird die Gesamtgröße aller (bzw. neu erfassten) PDBML-Dateien mit grauen (bzw. schwarzen) Balken dargestellt.

([Chruszcz08], [Ilari08], [Wlodawer08]) mit 89,5% und durch die Kernspinresonanz (NMR spectroscopy; [Wüthrich90], [Nietlispach04], [Billeter08]) mit 8,8%. Alle übrigen elf Verfahren machen zusammen fast 2% aus und nehmen daher eine untergeordnete Rolle ein. Die ersichtliche Ungleichverteilung ist später für die eigene Datenverwaltung aller Proteinstrukturen von Bedeutung ([Anders11]).

Die in der PDB archivierten Proteinstrukturen haben zwar meist hoch aufgelöste 3D-Atomkoordinaten. Die Atomkoordinaten als numerische Werte beschreiben jedoch 'nur' die Lage von Atomen im 3D Raum, sind als solche also nicht unbedingt von biologischer Bedeutung ([Gerstein00]). Folglich ist der Nutzen dieser PDB-Daten (zunächst) eingeschränkt, solange diese bspw. nicht mit anderen Daten(quellen) verknüpft werden ([Kasuya99], [Skolnick00]). Da die PDB primär ein Archiv für die Speicherung von Proteinstrukturen ist, gibt es hunderte anderer Projekte, die auf den PDB-Daten aufbauen und diese verschiedenartig weiterverarbeiten ([Berman08]), s. Anhang 2.B.

Sowohl das Speichern von Molekularstrukturen in Datenbanken als auch das Verknüpfen dieser Daten mit anderen Datenquellen ist eine technische und wissenschaftliche Herausforderung in der strukturellen Bioinformatik, jedoch unverzichtbar für biologische Erkenntnisgewinne ([Altman04]). Die Informatik begegnete diesen Herausforderungen oft mit originellen und kreativen Methoden, wobei manche zu nennenswerten wissenschaftlichen Entdeckungen führten ([Landsman09]). Solche Methoden werden in der wissenschaftlichen Literatur nicht bzw. kaum kommuniziert, da die Beschreibung einer Datenbank inklusive darauf aufbauender Software als unwesentlich angesehen wird ([Landsman09]). Die biologische Forschung wird sich im Gegensatz zu früher dem jedoch nicht verschließen können ([Howe08]), denn sie wird zwecks Erkenntnisgewinnen auf Datenbanken wesentlich mehr angewiesen sein ([B06]).

2.3 Die Datenbank-basierte PDB2-Infrastruktur

2.3.1 Die PDB Markup Language (PDBML)

Ein wohldefiniertes (engl. well-defined) - d. h. einheitliches und leicht zu parsendes - Datenformat ermöglicht zum einen das beliebige Abfragen (querying) von Daten eines solchen Formats. Ein wohldefiniertes Datenformat bietet zum anderen auch ein enormes Potential durch vergleichsweise einfache(re) Verknüpfungen mit anderen wohldefinierten Datenformaten. Durch das Verknüpfen mehrerer bzw. Korrelieren von Daten aus diversen Datenquellen lassen sich unter Umständen umfassendere biologische Erkenntnisse gewinnen.

Die Einführung und die Verwendung von XML (Extensible Markup Language) in den Lebenswissenschaften hatte sowohl einen großen Einfluss und Erfolg (s. bspw. [Achard01], [Buttler02], [Shabo06]), als auch technische, ökonomische und soziale Auswirkungen ([Wang05], [Adler06]). Ein XML Schema erleichtert nicht nur die Wiederverwendung von Daten, sondern auch den Import und Export selbiger, sowie das Vergleichen, Kombinieren und Verstehen der Daten ([Lynch08]). Ein einschlägiges XML Schema für Proteinstrukturen ([Westbrook05]), PDBML (PDB Markup Language), besteht aus selbstbeschreibenden Elementen (z. B. `<PDBx:title>`) und Attributen (z. B. `entry_id`), s. Abb. 2.4. Dieses Datenformat beinhaltet zwar dieselben Daten wie das ursprüngliche PDB-Format, es ist aber entscheidend für den Aufbau einer eigenen PDB2-Infrastruktur zwecks Verwaltung und Analyse von Proteinstrukturen.

2.3.2 Kurzeinführung in die Datenbank-basierte PDB2-Infrastruktur

Die überwiegende Mehrheit an Informationen zu Proteinstrukturen erfordert eine möglichst umfassende, durchdachte und systematisch erstellte Datenbank ([Sali98]). Darüber hinaus sind Anwenderspezifische Softwareentwicklungen notwendig, um diese Strukturdaten zumindest in eine Datenbank

```

<PDBx:audit_authorCategory>
  <PDBx:audit_author name="Pisabarro, M.T."></PDBx:audit_author>
  <PDBx:audit_author name="SSerrano, L."></PDBx:audit_author>
  <PDBx:audit_author name="Wilmanns, M."></PDBx:audit_author>
</PDBx:audit_authorCategory>
<PDBx:structCategory>
  <PDBx:struct entry_id="1BBZ">
    <PDBx:title>CRYSTAL STRUCTURE OF THE ABL-SH3 DOMAIN COMPLEXED WITH ~
      A DESIGNED HIGH-AFFINITY PEPTIDE LIGAND: IMPLICATIONS ~
      FOR SH3-LIGAND INTERACTIONS</PDBx:title>
    <PDBx:pdxb_descriptor>ABL TYROSINE KINASE, PEPTIDE P41
    </PDBx:pdxb_descriptor>
  </PDBx:struct>
</PDBx:structCategory>
<PDBx:struct_keywordsCategory>
  <PDBx:struct_keywords entry_id="1BBZ">
    <PDBx:pdxb_keywords>COMPLEX (TRANSFERASE/PEPTIDE)</PDBx:pdxb_keywords>
    <PDBx:text>COMPLEX (TRANSFERASE/PEPTIDE), SIGNAL TRANSDUCTION, SH3 ~
      DOMAIN</PDBx:text>
  </PDBx:struct_keywords>
</PDBx:struct_keywordsCategory>
<PDBx:exptlCategory>
  <PDBx:exptl entry_id="1BBZ">
    <PDBx:method>X-RAY DIFFRACTION</PDBx:method>
    <PDBx:crystals_number>1</PDBx:crystals_number>
  </PDBx:exptl>
</PDBx:exptlCategory>

```

Abbildung 2.4: Das modernere PDBML-Format am Beispiel des PDB-Datensatzes 1BBZ (Auszug), s. auch Abb. 2.1.

zu integrieren. Dabei ist insbesondere das Verarbeiten solcher Daten und deren Integration ein fortlaufender Prozess ([Berman00a]) mit vielen Abwägungen ([Warr08]).

Für den Datenimport von Proteinstrukturen in die PDB2-Infrastruktur werden sämtliche PDBML-Dateien zuerst geparkt. Anschließend werden bestimmte PDBML-Daten extrahiert und in ein relationales (Datenbank-)Schema transformiert (s. Teil II auf Seite 33). Alle relationalen und PDBML-Daten werden in einer einzigen Datenbank gespeichert, wobei die nicht-transformierten PDBML-Daten automatisch vom Datenbankmanagementsystem in einen XML-Baum transformiert werden, s. Abb. 2.6 im Anhang. Als Datenbanksystem wurde die DB2 Enterprise Server Edition für Linux, Unix und Windows (LUW) ab Version 9.x von IBM eingesetzt. Seit dieser Version bietet DB2 für LUW native XML-Unterstützung an (genannt pureXML), s. Abb. 2.5. Damit vereinfacht sich das logische (und physische) Datenbankdesign der PDB2-Infrastruktur (zusammengesetzt aus PDB und DB2) entscheidend, denn sämtliche PDBML-Daten können in lediglich zwei Tabellen gespeichert werden ([Anders11]). Ein rein relationales Datenbankschema (Mine 2 RDB¹) unter Verwendung von PostgreSQL umfasst derzeit rund 400 Tabellen ([Kinjo17]).

¹ <http://pd bj.org/mine-rdb-docs>

Die DB2-Infrastruktur mit dem two-hybrid - relational und (pure)XML - Speichermodell verbindet zwei - ehemals eigenständige - Datenbankwelten: Die traditionelle (d. h. relationale) und die XML-basierte Datenbankwelt. Ohne die pureXML-Funktionalität würde ein rein relationales Datenbankschema aus hunderten Tabellen (s. o.) und insgesamt mehreren tausend Spalten bestehen, s. Abb. 2.7 im Anhang. Neben diesem entscheidenden Vorteil der kombinierten Speicherung von relationalen und XML-basierten Daten bietet DB2 noch weitere Vorteile ([Anders11]). Zugunsten der uneingeschränkten Flexibilität bei Datenbankabfragen mittels SQL auf Konsolenebene wurde auf eine grafische Weboberfläche verzichtet, auch wenn diese Kopplung üblich ([Achard01], [Stevens01], [Baxevanis06]) und bei Biologen sehr populär ist ([Howe08]). In diesem Zusammenhang werden im Anhang 2.B einige PDB-bezogene Projekte aufgelistet.

Abbildung 2.5: Das DB2 XML-relationale Speichermodell (Quelle: [Bhambhri05]).

2.3.3 Zusammenfassung und Ausblick

Die entscheidenden Grundlagen für die PDB2-Infrastruktur waren sowohl die (syntaktischen) XML-Eigenschaften als auch die pureXML-Funktionalität von DB2. Das technische Design der Datenbank-/DB2-basierten PDB2-Infrastruktur war ein iterativer Prozess. Zahlreiche strukturebiologische Eigenschaften wurden sukzessive in das physische Datenbankdesign überführt. Ausführliche Tests und die Kenntnis von typischen, d. h. (struktur-)biologisch motivierten Datenbankabfragen führten zu weiteren Verbesserungen im physischen Datenbankdesign. Insgesamt sind damit vielseitige und vergleichsweise rasante Datenbanksuchen - über den gesamten Datenbestand - möglich. Wie pragmatisch die PDB2-Infrastruktur jedenfalls für die Analyse von Proteinstrukturen sein kann, wurde unter Verwendung von SECSTREME am Beispiel von IL8 und PI3 demonstriert. Ob neuere Ansätze wie Big Data-Technologien adäquate Infrastrukturalternativen sind, war nicht Untersuchungsgegenstand dieser Dissertationsschrift.

Kapitelanhang

2.A Abbildungen

Abbildung 2.6: XML-Baum zu Abb. 2.4 auf Seite 24.

Abbildung 2.7: Unveröffentlichtes Schema einer PDB-Referenzdatenbank des PDB-Konsortiums (1990er Jahre).

2.B PDB-bezogene Projektübersicht (Auswahl)

Im Folgenden werden einige Projekte aufgelistet, die auf PDB-Daten aufbauen. Da diese Projekte sehr heterogen sind, wurden diese grob kategorisiert.

2.B.1 Projekte zur strukturellen Klassifizierung von Proteinen

Viele biologische Anwendungen benötigen eine Klassifizierung der Proteinstrukturen, welche die PDB selbst nicht vornimmt. Es existieren mehrere Datenbanken bzw. Datenkollektionen, die eine grundlegende Klassifizierung anbieten. Populäre Vertreter sind insbesondere SCOP ([Murzin95]) und CATH(-Gene3D) ([Orengo97], [Sillitoe15a], [Sillitoe15b], [Dawson17]), sowie FSSP ([Holm94], [Holm96], [Holm97]) oder HSSP ([Sander91], [Schneider96], [Schneider97], [Dodge98]).

2.B.2 Datenbanken bzw. Datenkollektionen von Proteinannotationen

Mit dem alljährlichen Wachstum an frei zugänglichen Datenbanken bzw. Datenkollektionen (s. u. a. [Baxevanis03], [Galperin04], [Galperin05], [Galperin06], [Galperin07], [Galperin08], [Galperin09], [Cochrane10], [Galperin11], [Galperin12], [FernandezSuarez13], [FernandezSuarez14], [Galperin15], [Rigden16], [Galperin17], [Rigden18]) erhöhte sich in den letzten Jahren auch die Anzahl an Datenkollektionen, welche Proteine annotieren. Als Beispiele seien BioMolQuest ([Bukhman01]), iProtClass ([Wu01], [Huang03]), Columba ([Rother04], [Trissl05]), Atlas ([Shah05]), iProt([Jaya07]) genannt. Diese bestehen jeweils aus Daten, welche aus verschiedenen Datenquellen zusammengefügt und integriert wurden. Diese Datenquellen sind die PDB (exklusive 3D-Daten), SCOP, GO ([Ashburner00]) oder SWISS-PROT ([Bairoch00]).

2.B.3 Diverse Projekte des PDB-Konsortiums

Das PDB-Konsortium stellte immer wieder diverse Software zur Verfügung. Als Beispiele seien die WPDB ([Shindyalov95]), die PDBshell, die Macromolecular query language ([Shindyalov94]) und die PDBlib ([Chang94]) genannt, oder das mittlerweile veraltete OpenMMS toolkit ([Greer02]). Andere ([Shindyalov97], [Bourne03]) oder nicht publizierte Software des PDB-Konsortiums wird anderweitig aufgelistet ².

Abgesehen von der Softwareentwicklung gab es noch weitere Projekte. Die Entwicklung der PDB von einem Repository hin zu einem anspruchsvollen und komplexen wissensbasierten System ([Sussmann99]) wird durch folgende Projekte charakterisiert: Die ersten Schritte gab es in den frühen 1990er Jahren mit dem Aufsetzen einer winzigen relationalen Datenbank (Sybase) und Daten der Nucleic Acid Database ([Berman92]) inklusive 3D-Strukturen. Danach folgten die MOOSE-Datenbank, eine temporäre PDB+-Datenbank für die qualitative Evaluation von erhaltenen experimentellen Daten ([Weissig98], [Sussmann98]), sowie eine relationale Datenbank ([Bourne03], [Deshpande05]).

Darüber hinaus wurden durch das PDB-Konsortium auch verschiedene Visualisierungswerkzeuge ([Stampf95], [Peitsch95], [Rose11], [Quinn15]) entwickelt. Ein weiteres, längerfristig angelegtes Projekt war die "Sanierung" der PDB von Dateninkonsistenzen und der damit verbundenen Verbesserung der Datenqualität ([Bhat01], [Berman07], [Henrick08]). Die Protein Data Bank Japan entwickelte später einige Anwendungen für strukturbiochemische und biomedizinische Studien, wobei das PDBML Schema intensiv benutzt wurde ([Standley08]).

²<http://sw-tools.pdb.org>

2.C Nicht PDB-bezogene Projekte (Auswahl)

Neben den o. g. Projekten gab es darüber hinaus auch zahlreiche andere Projekte. Diese Projekte umfassten grob kategorisiert Abfragesysteme, Softwareentwicklungen oder auch Datenintegration.

2.C.1 Abfrage-/Speichersysteme

Spätestens mit der "Datenexplosion 1989-1992" ([Sussmann99], [Berman08]) brauchte das ehemals einfache PDB-Repository ein ergänzendes und moderneres Abfragesystem. Eines der ersten Projekte überhaupt war die Verwendung der logischen Programmiersprache Prolog ([Morffew86]). Zu den ersten Projekten unter Einbeziehung relationaler Datenbanksysteme gehörte(n)

- Projekte ([Todd84]) für die Visualisierung von Proteinstrukturen,
- das Birkbeck Integrated Protein Engineering Database (BIPED) Projekt ([Islam89]) unter Verwendung von Oracle für die Abfrage vorläufiger geometrischer Eigenschaften von Proteinen,
- das SESAM Projekt ([Huysmans91]) unter Verwendung von Sybase.

Die erste objektorientierte Datenbank für das Speichern und Abfragen von Proteinstrukturdaten wurde in den 1990er Jahren entwickelt ([Gray90]) unter Verwendung von Prolog oder der Abfragesprache Daplex und P/FDM als Datenbanksystem. Ein anderes Projekt war Omega ([Semertzidis93]), welches aus einer Zusammenstellung von strukturellen Informationen von Proteinen bestand.

2.C.2 Softwareentwicklungen

Die Strukturdaten der PDB werden üblicherweise im ursprünglichen PDB-Schema in Dateiform angeboten. Um diese Daten bspw. für Analysezwecke zugreifbar zu machen, mussten diese Dateien bzw. Daten erst geparkt und aufbereitet werden. Erst danach konnte auf diese aufbereiteten Daten anwendungsspezifisch zugegriffen werden, z. B. für Sequenz- oder Strukturüberlagerungen. Einige Softwareprojekte hatten sich zum Ziel gesetzt, diese sich wiederholende Datenvorverarbeitung zu systematisieren.

Als Vertreter von Open Source seien die Biochemical Algorithms Library ([Kohlbacher00]), die Bioinformatics Template Library ([Pitt01]), beide in C++ implementiert; BioJava ([Pocock00], [Holland08]); Bioperl ([Stajich02]), BioPython/PyPDB ([Hamelryck03], [Painter04], [Gilpin16]) oder HPDB-Haskell ([Gajda13]) genannt. Andere, eher weniger bekannte Entwicklungen sind in [Mangalam02] angegeben und bewertet.

2.C.3 Datenbank-theoretische Projekte

Weitere Arbeiten widmeten sich einem eher formalen und theoretischen Bereich im Datenbankwesen: Dem Design von Datenbankschemata, z. B. basierend auf

- Entity-Relationships wie das Hinzufügen von räumlichen Semantiken in 3D-Daten ([Ram04]) oder für die Sequenzintegration ([Bergholz01]),
- der Modellierung von Enhanced-Entity-Relationships in Bezug auf funktionelle Prozesse und 3D-Strukturen ([Elmasri07]),
- Diskussionen von konzeptuellen Modellierungsmethoden für biologische Daten ([Keet03]).

Implementierungs-orientierte Arbeiten mit dem Ziel Algorithmen als stored procedures (SP) oder user-defined functions (UDF) zu integrieren, um die Datenbankfunktionalität zu erhöhen, sind in ([Weh02]), [Chen03]) oder ([Hinneburg03]) beschrieben.

Teil II

Ausgewählte Publikationen

Managing the Protein Data Bank with DB2 pureXML

[Gerd Anders \(gerd.anders@ars.de\)](mailto:gerd.anders@ars.de)

Consultant and Bioinformatics Researcher
ARS Computer & Consulting

08 September 2011

[Matthias Nicola \(mnicola@us.ibm.com\)](mailto:mnicola@us.ibm.com)

Senior Software Engineer
IBM Silicon Valley Lab

The Protein Data Bank (PDB) is a unique worldwide repository of structural data about proteins. The PDB data is available in XML format in order to provide flexibility, extensibility, and ease of data exchange in the biological research community. Analyzing data in the PDB can help explain diseases, develop new drugs, or understand the interactions between different proteins. However, one of the key challenges is to efficiently store and query this information to find and extract information and correlations of interest. This article describes how to use the hybrid capabilities of DB2®— relational and pureXML® features — to manage and analyze PDB data.

Introduction

The Protein Data Bank (PDB.org) is a worldwide archive of structural data about biological molecules, mostly proteins. The Protein Data Bank (PDB) is managed by several member organizations responsible for depositing, maintaining, processing, and freely providing this biological data to the scientific community. To provide flexibility, extensibility, and ease of data exchange, the PDB data is available in XML format. This XML format is defined by an XML Schema known as the Protein Data Bank Markup Language (PDBML).

The structural information includes the 3-D coordinates of the atoms of the molecule(s) a protein consists of. These atomic coordinates are also referred to as the *3-D structure* or *tertiary structure*. The tertiary structure of a protein is closely coupled to its function. Thus, knowing the tertiary structure often helps in understanding the protein's intrinsic function. For example, the tertiary structure may be useful to explain diseases or develop new drugs. The tertiary structure can also be exploited to search the PDB for interactions between proteins.

The challenge

As of December 2010, the Protein Data Bank repository held 70,000 entries (XML documents) that contain more than 500 million atom coordinates. The total uncompressed size is more than 750 GB. Individual XML documents in the PDB range from a few MB to more than 1 GB in size.

Based on the rapid growth of the PDB repository in recent years ([Figure 1](#)), the size of the PDB is expected to continue to increase significantly. Consequently, searching and analyzing this information is becoming ever more challenging.

Figure 1. Growth of the PDB over the past 20 years

A typical approach to analyze PDB data is to write a custom application or a set of scripts that search the PDBML documents for the purpose of a very specific research question. The disadvantages of this approach include the facts that:

- Developing custom code each time new research is being conducted is very labor-intensive and time-consuming.
- The performance is often poor because all documents need to be parsed and searched, even if only a subset of them contain relevant information.
- It's often difficult to reuse or combine existing custom code to compose new or different queries against the PDB data.

DB2 V9.7.3 with pureXML was chosen to address these challenges, primarily because DB2 has the scalability and the XML capabilities required to process the expected volumes of PDBML documents. Additionally, DB2 is freely available for non-commercial usage via the IBM Academic Initiative. The goal was to store the PDB information in an efficient database schema, exploit relational and XML indexes for efficient search, and use XQuery and SQL/XML to express even complex queries against the PDB information.

The content of the Protein Data Bank

Before we discuss the DB2 database design for the PDB, it is helpful to understand the PDB data a little bit better.

The tertiary structure of a protein is experimentally determined (*solved*), predominantly by a method called *X-ray Diffraction* or *X-ray Crystallography*. Another less frequently used method is called *Solution NMR (Nuclear Magnetic Resonance)* or *NMR Spectroscopy*. The methods

for determining (solving) the protein structure lead to differences in how a protein structure is described in the generated XML documents, which is particularly reflected in the XML file sizes.

Proteins are dynamic molecules, which means that their tertiary structures may vary slightly, for example depending on their environment. Due to these variations, NMR methodically determines multiple instances (models) that represent slightly shifted tertiary structures for the same protein. Consequently, XML files with protein data produced by NMR can be very large in size, such as 100 MB to 1 GB, or more. Also, you will see later in this article how and why we use DB2 range partitioning to separate the first (*default*) model of a protein from its variations.

[Listing 1](#) shows an extract from one PDBML document. You can see four of 177 categories of information that can appear in such a document, including the authors of the study and the experimental method (<PDBx:exptlCategory>) used. The attribute entry_id represents the unique PDB identifier for this document.

Listing 1. Extract of a sample PDBML document (1BBZ.xml)

```
...
<PDBx:audit_authorCategory>
  <PDBx:audit_author pdbx_ordinal="1">
    <PDBx:name>Pisabarro, M.T.</PDBx:name>
  </PDBx:audit_author>
  ...
</PDBx:audit_authorCategory>
...
<PDBx:structCategory>
  <PDBx:struct entry_id="1BBZ">
    <PDBx:pdxb_descriptor>ABL TYROSINE KINASE, PEPTIDE P41
    </PDBx:pdxb_descriptor>
    <PDBx:title>CRYSTAL STRUCTURE OF THE ABL-SH3 DOMAIN COMPLEXED WITH
      A DESIGNED HIGH-AFFINITY PEPTIDE LIGAND: IMPLICATIONS FOR
      SH3-LIGAND INTERACTIONS
    </PDBx:title>
  </PDBx:struct>
</PDBx:structCategory>
...
<PDBx:struct_keywordsCategory>
  <PDBx:struct_keywords entry_id="1BBZ">
    <PDBx:pdxb_keywords>COMPLEX(TRANSFERASE/PEPTIDE)
    </PDBx:pdxb_keywords>
    <PDBx:text>COMPLEX (TRANSFERASE-PEPTIDE), SIGNAL TRANSDUCTION, SH3 DOMAIN,
      COMPLEX (TRANSFERASE-PEPTIDE) complex
    </PDBx:text>
  </PDBx:struct_keywords>
</PDBx:struct_keywordsCategory>
...
<PDBx:exptlCategory>
  <PDBx:exptl entry_id="1BBZ" method="X-RAY DIFFRACTION">
    <PDBx:crystals_number>1</PDBx:crystals_number>
  </PDBx:exptl>
</PDBx:exptlCategory>
...
```

The test database

Due to time and resource constraints, we decided to use only a subset of the total available PDB data volume to prototype and evaluate the storage, indexing, and querying of PDBML documents in a DB2 database. Therefore, a representative sample of 6,029 documents was selected,

which amounts to 83 GB and roughly 10 percent of the total volume of the PDBML archive as of December 2010. This set of documents contains approximately 1.7 billion XML elements, out of which approx. 1.54 billion elements describe tertiary protein structures through atom coordinates and other information.

A representative sample of PDBML documents must accurately reflect the ratio of documents with molecule information produced by X-ray Diffraction (smaller documents, 83 percent of all documents) vs. Solution NMR (larger documents, 16 percent of all documents). This ensures that the database configuration and queries are tested with a realistic mix of small and large documents.

The database server available for this study was a Sun X4600 M2 with eight dual-core processors (AMD Opteron 8220) and 256GB main memory. The operating system was Ubuntu 64-bit Linux®. The storage consisted of 10 hard drives (698 GB each; 7,200 rpm), organized as a single logical volume (RAID 5) using a hardware controller.

Database design recommendations for PDB

This section describes a set of database design recommendations that lead to simple and efficient database support for storing and analyzing PDB data. These recommendations address the database schema, the choice between XML and relational storage, definition of indexes, and physical data organization with partitioning and clustering options.

Hybrid XML/Relational storage

PDBML documents currently contain up to 177 categories of information, most of them optional. The large number of optional PDBML elements allow the documents to be very flexible and highly variable. A fully relational database schema would require hundreds of tables to represent PDBML. Such a relational database schema for the PDB was developed in 2005 and is shown in [Figure 2](#). With more than 400 tables and more than 3,000 columns, the complexity of this schema is overwhelming. It is extremely difficult to understand and query such a database schema because a single PDB entry is broken up and scattered over hundreds of tables, making it hard for users to know which information resides in which table. Therefore, keeping most of the PDBML information in its original XML format and storing it in a single XML column results in a much simpler database design and retains the data in a format that users naturally understand.

Figure 2. Diagram of a fully relational database schema for PDBML

One notable exception to the high variability of the PDBML data is the atom coordinates and their related labels, which follow a flat and regular structure repeated for every atom in a molecule, as illustrated in [Listing 2](#). Since proteins commonly consist of thousands or tens of thousands of atoms, atom coordinates often represent 90 percent or more of a PDBML document.

Listing 2. Atom coordinates in a PDBML document

```
<PDBx:atom_siteCategory>
  <PDBx:atom_site id="1">
    <PDBx:B_iso_or_equiv>37.41</PDBx:B_iso_or_equiv>
    <PDBx:Cartn_x>1.039</PDBx:Cartn_x>
    <PDBx:Cartn_y>16.834</PDBx:Cartn_y>
    <PDBx:Cartn_z>18.876</PDBx:Cartn_z>
    <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
    <PDBx:auth_atom_id>N</PDBx:auth_atom_id>
    <PDBx:auth_comp_id>ASN</PDBx:auth_comp_id>
    <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
    <PDBx:group_PDB>ATOM</PDBx:group_PDB>
    <PDBx:label_alt_id xsi:nil="true" />
    <PDBx:label_asym_id>A</PDBx:label_asym_id>
    <PDBx:label_atom_id>N</PDBx:label_atom_id>
    <PDBx:label_comp_id>ASN</PDBx:label_comp_id>
    <PDBx:label_entity_id>1</PDBx:label_entity_id>
    <PDBx:label_seq_id>1</PDBx:label_seq_id>
    <PDBx:occupancy>1.00</PDBx:occupancy>
    <PDBx:pdex_PDB_model_num>1</PDBx:pdex_PDB_model_num>
    <PDBx:type_symbol>N</PDBx:type_symbol>
  </PDBx:atom_site>
  <PDBx:atom_site id="2">
    <PDBx:B_iso_or_equiv>36.15</PDBx:B_iso_or_equiv>
    <PDBx:Cartn_x>-0.213</PDBx:Cartn_x>
    <PDBx:Cartn_y>16.205</PDBx:Cartn_y>
    <PDBx:Cartn_z>18.364</PDBx:Cartn_z>
    ...
  </PDBx:atom_site>
  <PDBx:atom_site id="3">
    <PDBx:B_iso_or_equiv>33.97</PDBx:B_iso_or_equiv>
    <PDBx:Cartn_x>-0.549</PDBx:Cartn_x>
    <PDBx:Cartn_y>16.779</PDBx:Cartn_y>
    <PDBx:Cartn_z>16.986</PDBx:Cartn_z>
    ...
  </PDBx:atom_site>
```

```
...
</PDBx:atom_siteCategory>
```

The flat and regular structure of the atom information makes a perfect fit for traditional relational tables. In fact, the atom coordinates and labels are non-hierarchical data for which XML is not the best choice. Therefore, we decide on a hybrid database schema that stores the *atom_site* information in a relational table and the remainder of each PDBML document in an XML column, but with the `<atom_siteCategory>` removed from the document. This has several advantages:

- The reduced PDBML documents are much smaller, which improves insert and load performance, as well as XML query performance. The XML parsing effort upon insert or load is reduced by approximately 90 percent.
- The atom information takes less space in relational columns than in their verbose XML representation.
- The atom data can be queried with traditional relational methods, which for non-hierarchical data is more efficient than XML navigation.
- Since each atom is represented in a separated row, indices can help speed up the search for specific atoms within a given PDBML entry.

The chosen database schema consists of two tables, shown in [Listing 3](#). The first (`xmlrpdb.pdbxml`) has one row for each PDB entry. This table has only two columns:

- The primary key `pdb_id` holds the four-character PDB entry identifier from the XML attribute `entry_id`.
- The XML column `pdbxml_file` holds the entire PDBML document except the `<atom_siteCategory>`.

The second table (`xmlrpdb.atom_site`) contains one relational row for each atom coordinate (i.e., for each `<atom_site>` element in a PDBML document). The column `pdb_id` is the foreign key that links atom coordinates to the corresponding PDBML document in the `pdbxml` table.

Both tables are stored in table spaces with a 32-KB page size to maximize the performance analytical queries that read large numbers of rows.

Listing 3. Hybrid XML/relational database schema for PDB in DB2

```
CREATE TABLE xmlrpdb.pdbxml (
  pdb_id          CHAR(4) NOT NULL,
  pdbxml_file     XML NOT NULL,
  PRIMARY KEY (PDB_ID))
IN ts_data32k INDEX IN ts_index32k;

CREATE TABLE xmlrpdb.atom_site (
  pdb_id          CHAR(4) NOT NULL,
  atom_site_id    INTEGER NOT NULL,
  auth_asym_id    VARCHAR(10) WITH DEFAULT NULL,
  auth_atom_id    VARCHAR(20) NOT NULL,
  auth_comp_id    VARCHAR(3) NOT NULL,
  auth_seq_id     VARCHAR(20) NOT NULL,
  b_iso_or_equiv  DECIMAL(7,3) NOT NULL,
  b_iso_or_equiv_esd DECIMAL(7,3) WITH DEFAULT NULL,
  cartn_x         DECIMAL(7,3) NOT NULL,
  cartn_x_esd     DECIMAL(7,3) WITH DEFAULT NULL,
  cartn_y         DECIMAL(7,3) NOT NULL,
```

```

cartn_y_esd          DECIMAL(7,3) WITH DEFAULT NULL,
cartn_z              DECIMAL(7,3) NOT NULL,
cartn_z_esd          DECIMAL(7,3) WITH DEFAULT NULL,
group_pdb            VARCHAR(10) NOT NULL,
label_alt_id         VARCHAR(10) WITH DEFAULT NULL,
label_asym_id        VARCHAR(10) WITH DEFAULT NULL,
label_atom_id        VARCHAR(20) WITH DEFAULT NULL,
label_comp_id        VARCHAR(10) NOT NULL,
label_entity_id      SMALLINT NOT NULL,
label_seq_id         SMALLINT WITH DEFAULT NULL,
occupancy            DECIMAL(7,3) NOT NULL,
occupancy_esd        DECIMAL(7,3) WITH DEFAULT NULL,
pdbx_pdb_atom_name   VARCHAR(10) WITH DEFAULT NULL,
pdbx_pdb_ins_code    VARCHAR(10) WITH DEFAULT NULL,
pdbx_PDB_model_num SMALLINT NOT NULL,
type_symbol          VARCHAR(10) WITH DEFAULT NULL,
PRIMARY KEY (pdb_id, atom_site_id),
FOREIGN KEY (pdb_id) REFERENCES xmlrpdb.pdbxml(pdb_id),
CONSTRAINT group_chk CHECK (group_PDB in ('ATOM', 'HETATM'))
) IN ts_atom_data_32k INDEX IN ts_atom_index32k;

```

Optionally, CHECK constraints can be defined on the `pdbxml` table to ensure that the four-character PDB identifier conforms to the PDB standard. The first character must be a number between 1 and 9, and the next three characters must be a number between 0 and 9 or an uppercase character between A and Z (see [Listing 4](#)).

Listing 4. CHECK constraints to enforce proper `pdb_id` values

```

ALTER TABLE xmlrpdb.pdbxml
  ADD CHECK (SUBSTR(pdb_id, 1, 1) BETWEEN '1' AND '9')
  ADD CHECK ((SUBSTR(pdb_id, 2, 1) BETWEEN '0' AND '9') OR
             (SUBSTR(pdb_id, 2, 1) BETWEEN 'A' AND 'Z'))
  ADD CHECK ((SUBSTR(pdb_id, 3, 1) BETWEEN '0' AND '9') OR
             (SUBSTR(pdb_id, 3, 1) BETWEEN 'A' AND 'Z'))
  ADD CHECK ((SUBSTR(pdb_id, 4, 1) BETWEEN '0' AND '9') OR
             (SUBSTR(pdb_id, 4, 1) BETWEEN 'A' AND 'Z'));

```

Populating the hybrid database schema

The conceptual process of inserting a PDBML document into our hybrid database schema is illustrated in [Figure 3](#). The `<atom_siteCategory>` data needs to be extracted and removed from the XML document, and inserted into the relational `atom_site` table (blue). The reduced document itself is inserted into the `pdbxml` table. We call this process *atom site separation*.

Figure 3. Hybrid storage of a PDBML document with separation of `atom_site` data

Due to high data volume, the atom site separation (the population of the hybrid database schema) needs to have high performance. Hence, costly XML parsing should be reduced as much as possible. Revisiting the atom coordinates in XML format in [Listing 2](#), we find that 94.5 percent of the characters are markup, and only 5.5 percent of the characters are actual values. Hence, the ratio of markup to value is extremely high, which means that a lot of XML parsing may be required to extract a comparatively small amount of usable data. You will understand shortly how this consideration has affected our decision of how to populate the two tables.

One option to populate the relational `atom_site` table is to use `INSERT` statements with an `XMLTABLE` function. Such a statement can parse the entire PDBML document and extract the atom information to insert as relational rows. Additionally, an XQuery Update expression can delete the `<atom_siteCategory>` subtree from the each PDBML document inserted into the `pdxml` table. Such an XQuery Update expression can also be part of an `INSERT` statement so that the `<atom_siteCategory>` is removed before writing it into the XML column, rather than performing a separate step after the insert.

Another option is to use a special-purpose preprocessor outside the database to extract the atom data into a relational flat file and remove it from each PDBML document. Such a preprocessor was implemented in C++, and it has the following benefits:

- It can add desirable annotations to the data, such as information from sequence and structure alignments or application-dependent geometric transformations like rotations or translations of atomic coordinates.
- It can be implemented without using a general-purpose XML parser. Instead, it is designed and optimized for the specific file structure of PDBML documents. It exploits special

knowledge about the flat structure of the atom data, the existence of newline characters between elements, and other characteristics. As a result, the specialized preprocessor is at least 10 times faster than any solution with XML parsing.

Preprocessing the data set of 6,029 gzipped PDBML documents (i.e., 83 GB unzipped) and loading the prepared data into the `pdbxml` table and `atom_site` table took only 1 hour and 44 minutes. The preprocessor is available for download (see [Download](#)).

Compression

Considering the data volume in the PDB archive as well as its rapid growth, it is useful to compress the data in DB2. This reduces the storage consumption and improves performance. Although compression and decompression in DB2 consumes some additional CPU cycles, compression also reduces the number of physical I/O operations necessary to read a certain amount of data from disk. Furthermore, compressed pages of a DB2 table space remain compressed in the DB2 buffer pool in main memory. As a result, compression allows more data to be in memory than without compression, which increases the buffer-pool hit ratio and makes higher utilization of the available memory. We found that the I/O and memory benefits of compression outweigh the additional CPU cost and lead to overall higher performance.

The following commands in [Listing 5](#) were used to compress both tables.

Listing 5. Enabling compression and REORG tables

```
ALTER TABLE xmlrpdb.pdbxml COMPRESS YES;  
REORG TABLE xmlrpdb.pdbxml LONGLOBDATA RESETDICTIONARY;  
  
ALTER TABLE xmlrpdb.atom_site COMPRESS YES;  
REORG TABLE xmlrpdb.atom_site LONGLOBDATA RESETDICTIONARY;
```

The reduction in space consumption is summarized in [Table 1](#). After compression, the information contained in the 6,029 PDBML documents can be stored in 67.4 percent fewer pages (i.e., three times less space than without compression).

Table 1. Space savings achieved by compression

	Before compression	After compression	Savings
xmlrpdb.pdbxml	176,256 pages	44,736 pages	74.6 percent
xmlrpdb.atom_site	264,960 pages	99,264 pages	62.5 percent
Total	441,216 pages	144,000 pages	67.4 percent

With a page size of 32 KB, the final storage of 144,000 pages is equivalent to 4.4 GB, which is only 5.3 percent of the original raw data volume of 83 GB. If we extrapolate this ratio to the total current size of the PDB archive, we find that the 0.75 TB of PDB information would be stored in DB2 using only approximately 40.7 GB of space, plus some space for indices.

This tremendous saving of storage stems from two factors. First, the high ratio of markup to value in the atom information is eliminated by converting the atom coordinates to relation format in the

preprocessing step. Second, DB2 compression shrinks the remaining XML and relational data by a factor of 3.

Database partitioning

Despite the significant reduction in space consumption, the PDB data volume continues to grow fast. Also, the response time of complex analytical queries can be reduced by spreading the data across multiple database partitions, such that all partitions work on their assigned data in parallel. These database partitions can reside on the same machine to exploit all the CPU power of a multi-core system, or they can be spread across multiple machines in a shared-nothing configuration. The DB2 Database Partitioning Feature (DPF) is available through IBM InfoSphere® Warehouse, a software package that contains DB2 with advanced features, as well as additional design, reporting, and database management tools.

Using the DPF, we recommend to distribute the data in the `pdbxml` table and `atom_site` table across the database partitions by hashing on the values of the `pdb_id` column. This is achieved by adding the clause `DISTRIBUTE BY HASH(pdb_id)` to the respective `CREATE TABLE` statement. The large number of distinct values in the `pdb_id` column ensures a relatively even distribution of rows over the database partitions. Distributing both tables by hashing on their join key (`pdb_id`) also ensures that all `atom` rows for a given PDBML document are stored in the same database partition as the PDBML document itself. This collocation implies that joins between the two tables can always be evaluated within each of the database partitions and never require data to be shipped across partitions.

Range partitioning

Range partitioning (also known as *table partitioning*) enables you to partition the data in a table according to the value in a specified column, such that rows with the same value will reside in the same partition. The concept of range partitioning is orthogonal to the database partitioning. If database partitioning and range partitioning are used together, the rows in a table are first hashed across the database partitions and then range-partitioned within each database partition.

Range partitioning can serve multiple purposes. One purpose is easier roll-in and roll-out of new and old data, respectively. Another purpose is to improve performance based on *partition elimination* when the DB2 query optimizer determines that only a subset of the partitions need to be examined to answer a particular query. For the PDB, range partitioning was deployed to benefit from partition elimination, rather than to simplify roll-in and roll-out of data.

We decided to range-partition the `atom_site` table by the `pdbx_PDB_model_num` column for the following reason: Remember that the tertiary structure of a protein can be experimentally determined with a method called NMR, which produces multiple tertiary structures for the same protein. These variations are called *models* and are numbered by the field `pdbx_PDB_model_num`. A value of `pdbx_PDB_model_num = 1` identifies the first (default) model of a protein. The additional variations are the non-default models of the same protein and have `pdbx_PDB_model_num >= 2`. Proteins that have been structurally determined by X-ray Diffraction have only one model with `pdbx_PDB_model_num = 1`.

[Listing 6](#) shows the extended definition of the `atom_site` table with range partitioning. All atom coordinates that belong to the first model (`pdxb_PDB_model_num = 1`) are stored in one partition, whereas any variations (`pdxb_PDB_model_num >= 2`) are stored in another. Although only about 16 percent of all proteins currently in the PDB have variations produced by NMR, the number of their variations is so large that both partitions have roughly the same number of records.

Listing 6. Table definition with range partitioning

```
CREATE TABLE xmlrpdb.atom_site (
  pdb_id          CHAR(4) NOT NULL,
  ...
  pdbx_PDB_model_num SMALLINT NOT NULL,
  type_symbol      VARCHAR(10) WITH DEFAULT NULL,
  PRIMARY KEY (pdb_id, atom_site_id),
  FOREIGN KEY (pdb_id) REFERENCES xmlrpdb.pdbxml(pdb_id),
  CONSTRAINT group_chk CHECK (group_PDB in ('ATOM', 'HETATM'))
)
-- IN ts_atom_data_32k
INDEX IN ts_atom_index32k
PARTITION BY RANGE (pdxb_PDB_model_num)
(
  PARTITION DEF_MODELS      STARTING (1) ENDING (1) IN TS_ATOM_DATA1_32K,
  PARTITION NON_DEF_MODELS  STARTING (2) ENDING (MAXVALUE) IN TS_ATOM_DATA2_32K
);
```

We have chosen this range-partitioning scheme because many PDB queries typically differentiate between default and non-default protein models, and can therefore benefit from the partitioning. For example, a query that analyzes all or most of the default models only needs to scan the partition `DEF_MODELS`, which reduces the required I/O by half.

Multi-dimensional clustering

In addition to range-partitioning, *multi-dimensional clustering* (MDC) can be used to cluster the rows in a table based on one or more columns. Rows that have the same value in the clustering columns are physically stored together in the same storage cell. This can greatly improve the performance of queries that constrain and select data along one or multiple clustering dimensions. Like the DPF, MDC is also available through IBM InfoSphere Warehouse.

The choice of clustering columns needs to be based on the expected query workload so the clustering supports the most common and most critical queries. For example, many PDB queries might search the atom data based on the amino acid involved. Therefore, it can be beneficial to cluster the `atom_site` table based on the column `label_comp_id`, which, in most documents, contains the three-letter code for the amino acid. To achieve this clustering, add the following clause to the second `CREATE TABLE` statement in [Listing 3](#): `ORGANIZE BY DIMENSIONS(label_comp_id)`.

Another example is to cluster the `atom_site` table by the `group_PDB` column. We have evaluated this clustering for several sample queries that restrict their search to a single `group_PDB` value (i.e., "HETATOM") and found that it can improve query performance fourfold.

PDB queries and performance

In this section, we discuss two sample queries to demonstrate:

- The ease with which even complex analysis of PDB data can be carried out.
- The benefit of the database design decisions described in the previous sections.

The query in [Listing 7](#) selects the PDB identifier, the resolution, and the description from all PDB entries where the experimental method is "X-RAY DIFFRACTION" and the resolution (`ls_d_res_high`) is less than 2.5. The resolution is expressed in Ångström (1Å = 0.1 nanometer) and serves as a quality metric for the analysis of the atom structures. Structures with a resolution less than 2Å are high-resolution structures (i.e., the locations of their atoms could be determined very accurately). Structures with a resolution greater than 3Å are less accurate and usually ignored.

Listing 7. Query the top 10 records with the best X-ray resolution

```
SELECT pdb_id, x.resolution, x.pdbx_descriptor
FROM xmlrpdb.pdbxml,
  XMLTABLE
    ('$PDBXML_FILE/*:datablock/*:refineCategory/*:refine[
      @pdbx_refine_id = "X-RAY DIFFRACTION" and *:ls_d_res_high <= 2.5 ]'
  COLUMNS
    resolution      DEC(9,5)      PATH '*:ls_d_res_high',
    pdbx_descriptor VARCHAR(2000) PATH '.../*:structCategory/*:struct/*:pdbx_descriptor'
  ) AS x
-- WHERE
--   upper(x.pdbx_descriptor) LIKE '%UNKNOWN%' or
--   upper(x.pdbx_descriptor) LIKE '%UNCHARACTERIZED%'
ORDER BY x.resolution
FETCH FIRST 10 ROWS ONLY;
```

The result of this query is shown in [Listing 8](#). One of the benefits of using DB2 pureXML as opposed to custom code is that it's easy to modify SQL/XML queries to refine the search. For example, [Listing 7](#) contains three comment lines with an additional `WHERE` clause. They can be used to further filter the descriptor to find those structures that are not or could not be characterized yet.

Listing 8. Result produced by the query in Listing 7

PDB_ID	RESOLUTION	PDBX_DESCRIPTOR
2VB1	0.65000	LYSOZYME C (E.C.3.2.1.17)
2B97	0.75000	Hydrophobin II
2OV0	0.75000	PROTEIN
2I16	0.81000	Aldose reductase (E.C.1.1.1.21)
2I17	0.81000	Aldose reductase (E.C.1.1.1.21)
2HS1	0.84000	HIV-1 Protease V32I mutant (EC 3.4.23.16)
2F01	0.85000	Streptavidin
2OL9	0.85000	SNQNNF peptide from human prion residues 170-175
2PF8	0.85000	Aldose reductase (E.C.1.1.1.21)
2P74	0.88000	Beta-lactamase CTX-M-9a (E.C.3.5.2.6)

10 record(s) selected.

The predicates in the query in [Listing 7](#) are weakly selective so that a full table scan of the `pdbxml` table is required. [Table 2](#) summarizes how the performance of this table scan query has benefited from two of our design decisions: atom site separation and compression. In our environment, this table scan was I/O-bound. DB2 compression mitigated the I/O bottleneck and reduced the query elapsed time by more than 40 percent (from 244 to 128 seconds). Extracting the atom site data

into a separate relational table greatly reduced the size of the `pdboxml` table, improving the query performance by almost 4 1/2 times, from 138 to 31 seconds.

Table 2. Response times (without indices) of the query in Listing 7

Atom site separation	Compression	Response time
No	No	244 seconds
No	Yes	138 seconds
Yes	Yes	31 seconds

[Listing 9](#) shows another sample query, which determines how often different atoms — or ions — occur in different compounds. The `WHERE` clause restricts the search to so-called *hetero* atoms and only considers the first model of each protein.

Listing 9. Analysis of hetero atom occurrences

```
SELECT label_atom_id          AS "Atom",
       COALESCE(label_comp_id, 'none') AS "Compound",
       COUNT(*)               AS "Occurrences"
FROM xmlrpdb.atom_site
WHERE group_PDB = 'HETATM'
      AND pdbox_PDB_model_num = 1
GROUP BY label_atom_id, label_comp_id
ORDER BY COUNT(*), label_comp_id DESC;
```

A subset of the result rows is shown in [Listing 10](#). The most frequently detected chemical compound is water (HOH) with oxygen (O) as one of its atoms. The reported number of hydrogens, denoted by H1 and H2 for HOH, is low because detecting hydrogens requires a very high resolution that is not always achieved.

(Human) hemoglobin is a protein consisting of multiple molecules, and such a molecule can interact with a non-protein compound called *heme*. A heme (HEM) is a multi-atom, non-proteinaceous organic structure capable of positioning an iron (FE) ion in its center. This iron ion, on its hand, is critical for oxygen binding. The result in [Listing 10](#) shows that iron occurs frequently together with heme compounds. Although this is a simple example, it demonstrates how efficient it has become to detect meaningful correlations in the PDB data and to gain better understanding of how proteins function and interact on a molecular level.

Listing 10. Subset of the result produced by the query in Listing 9

Atom	Compound	Occurrences
O	HOH	1571965
MG	MG	7159
...		
H1	HOH	1858
H2	HOH	1858
ZN	ZN	1664
...		
CL	CL	1318
CA	CA	1295
...		
FE	HEM	379
NA	HEM	379

[Table 3](#) shows how our database design choices for atom site separation, compression, range partitioning, and multi-dimensional clustering can provide excellent performance, even when no query-specific indexes exist.

Table 3. Response times (without indices) of the query in Listing 9

Atom site separation	Compression	Range partitioning	MDC	Response time
Yes	Yes	No	No	38.7 seconds
Yes	Yes	Yes	No	25.8 seconds
Yes	Yes	Yes	Yes	5.5 seconds

Summary

This article has described how to use pureXML and relational data management features in DB2 to efficiently store and query the Protein Data Bank (PDB). Based on the intrinsic characteristics of the protein data, we have designed an optimized hybrid database schema. For best performance and minimal space consumption, we recommend using database partitioning, range partitioning, compression, and multi-dimensional clustering. Additionally, a combination of XML indices and relational indexes can further improve query performance. The DB2-based PDB continues to be used for investigations, such as searching the entire PDB for certain protein interactions and to help explain unusual interactions on the structural level.

Acknowledgements

The development of the DB2-based PDB was done in the Structural Bioinformatics research group of Maria Teresa Pisabarro, Biotechnology Center, Technical University Dresden, Germany. The project was financed by a scholarship from the foundation of SAP co-founder Klaus Tschira. Also, thanks to Henrik Loeser for his help with the work described in this article and the Berlin Institute of Medical Systems Biology (BIMSB) of the Max Delbrück Center (MDC) for Molecular Medicine Berlin-Buch, Germany, for providing the production server.

Downloads

Description	Name	Size
C++-preprocessor and a few DB2 sample queries	db2pdb_download.zip	965KB

Resources

Learn

- Visit PDB.org for more information about the Protein Data Bank and the XML format PDBML.
- Learn more about the [XML format PDBML](#).
- For an introduction to PDB data, read "[Understanding PDB Data: Looking at Structures](#)."
- Get the full [XML document from which extracts are shown in Listings 1 and 2](#).
- Gain comprehensive knowledge of DB2 pureXML with the [DB2 pureXML Cookbook](#).
- Read more about compression, partitioning, and clustering of XML data in the article "[Enhance business insight and scalability of XML data with new DB2 9.7 pureXML features](#)."
- For more DB2 pureXML resources, explore the [DB2 pureXML enablement wiki](#).
- Read the [Native XML Database blog](#) to stay up to date on latest news and XML tips and tricks.
- Learn about the [IBM Academic Initiative](#).
- Learn more about Information Management at the [developerWorks Information Management zone](#). Find technical documentation, how-to articles, education, downloads, product information, and more.
- Stay current with [developerWorks technical events and webcasts](#).
- Follow [developerWorks on Twitter](#).

Get products and technologies

- Download [DB2 Express-C](#) for free.
- Build your next development project with [IBM trial software](#), available for download directly from developerWorks.

Discuss

- [Participate in the discussion forum for this content](#).
- Check out the [developerWorks blogs](#) and get involved in the [developerWorks community](#).

About the authors

Gerd Anders

Gerd Anders holds a degree in computer science from the Humboldt University Berlin. He focuses on bio-informatics and the use of database systems to efficiently manage and process very large amounts of data in and for the life sciences. The DB2-based PDB is an essential project of his Ph.D. thesis, providing a valuable basis for several DB2-driven applications and exploiting the entire potential of the PDB. In addition, he was also a tester of the closed-beta version of DB2 9.5 for Linux, UNIX, and Windows.

Matthias Nicola

Matthias Nicola is a Senior Technical Staff Member at IBM's Silicon Valley Lab in San Jose, CA. He focuses on DB2 performance and benchmarking, XML, temporal data management, in-database analytics, and other emerging technologies. He also works closely with customers and business partners to help them design, optimize, and implement DB2 solutions. Previously, Matthias worked on data warehouse performance at Informix Software. He received his Ph.D. in computer science from the Technical University of Aachen, Germany.

© Copyright IBM Corporation 2011

(www.ibm.com/legal/copytrade.shtml)

Trademarks

(www.ibm.com/developerworks/ibm/trademarks/)

The Intrinsic Pepsin Resistance of Interleukin-8 Can Be Explained from a Combined Bioinformatical and Experimental Approach

Gerd Anders, Ulrich Hassiepen, Stephan Theisgen, Stephan Heymann, Lionel Muller, Tania Panigada, Daniel Huster, and Sergey A. Samsonov

Abstract—Interleukin-8 (IL-8, CXCL8) is a neutrophil chemotactic factor belonging to the family of chemokines. IL-8 was shown to resist pepsin cleavage displaying its high resistance to this protease. However, the molecular mechanisms underlying this resistance are not fully understood. Using our in-house database containing the data on three-dimensional arrangements of secondary structure elements from the whole Protein Data Bank, we found a striking structural similarity between IL-8 and pepsin inhibitor-3. Such similarity could play a key role in understanding IL-8 resistance to the protease pepsin. To support this hypothesis, we applied pepsin assays confirming that intact IL-8 is not degraded by pepsin in comparison to IL-8 in a denaturated state. Applying ^1H - ^{15}N Heteronuclear Single Quantum Coherence NMR measurements, we determined the putative regions at IL-8 that are potentially responsible for interactions with the pepsin. The results obtained in this work contribute to the understanding of the resistance of IL-8 to pepsin proteolysis in terms of its structural properties.

Index Terms—Protein-protein interactions, computational biology, pepsin resistance, pepsin inhibitor, interleukin-8

1 INTRODUCTION

IL-8 belongs to a family of chemokines, which consist of small chemotactic cytokines. It was first identified in 1987 as a novel type of neutrophil-activating cytokine [1]. Human IL-8 is 77 amino acids long and found in both monomeric and dimeric forms [2]. The human fetus as well as newborns take up biologically significant quantities of IL-8 in amniotic fluid and breast milk. It was previously reported that IL-8 is highly resistant to pepsin, an aspartic proteinase [3]. In particular, Maheshwari et al. applied immunological analysis, Western blot techniques, cell proliferation and chemotaxis assays to demonstrate this resistance. Using molecular docking simulations they also proposed a molecular model, which explained the resistance by absence of structural complementarity between IL-8 and pepsin binding site: potential cleavage sites for protease on IL-8 are not accessible for pepsin

when disulfide bridges in IL-8 are intact [3]. There are not many characterized proteins that show a pepsin resistance or inhibit pepsin [4]. *Ascaris suum* (*A. suum*) produces pepsin-specific inhibitors [5], [6], [7], e.g., the 149-residue pepsin inhibitor-3 (PI-3; [8], which has no significant sequence similarity with any other proteins from other organisms [8], [9]. PI-3 is one of a few structurally determined protein inhibitors of aspartic proteinases [10], and it reversibly inhibits the proteolytic activities of the gastric enzymes pepsin and gastricsin [11]. PI-3 as pepsin inhibitor was also shown to affect immune responses [12] suggesting an immunomodulatory function [13]. It has been suggested that the inhibitor from *A. suum* may function to protect the nematode during passage through the host's digestive system or to evade immune responses [14], [12], some causing production of IL-8 [15], [16], [17], [18]. A direct interaction between an *A. suum*-derived chemotactic factor and the host's chemotactic network, via a G protein-coupled receptor, which interacts with the IL-8 receptor pathway, is evident [19].

Although IL-8 resistance to pepsin gained attention before, previous studies did not explain the relevant underlying molecular mechanisms. In the *in silico* part of this study, we aimed at understanding this resistance by analyzing structural properties of IL-8 such as three-dimensional (3D) arrangements of its secondary structure elements (SSEs) using a combination of geometric hashing and clique detection implemented in a in-house database containing the data from the whole Protein Data Bank (PDB). Geometric hashing is a highly efficient technique for matching geometric features within a database [20]. The particular strength of geometric hashing in structural biology is its ability to find highly similar 3D substructures in proteins [21], [22], [23]. It was previously successfully applied for

- G. Anders is with ARS Computer and Consulting GmbH, Munich 80339, Germany, and the Institute of Medical Physics and Biophysics, University of Leipzig, Leipzig 04107, Germany. E-mail: gerd.anders@ars.de.
- U. Hassiepen and L. Muller are with the Novartis Institutes for BioMedical Research (NIBR), Center for Proteomic Chemistry, Novartis Pharma AG, Basel, Switzerland. E-mail: {ulrich.hassiepen, lionel.muller}@novartis.com.
- S. Theisgen and D. Huster are with the Institute of Medical Physics and Biophysics, University of Leipzig, Leipzig 04107, Germany. E-mail: stheisgen@gmx.de, daniel.huster@medizin.uni-leipzig.de.
- S. Heymann is with ICP Healthcare GmbH Berlin, Berlin 13125, Germany. E-mail: heyman@informatik.hu-berlin.de.
- T. Panigada is with the Novartis Institute for BioMedical Research (NIBR), Analytical Sciences and Imaging, Novartis Pharma AG, Basel, Switzerland. E-mail: tania.panigada@novartis.com.
- S.A. Samsonov is with the Structural Bioinformatics, BIOTEC TU Dresden, Dresden 01307, Germany. E-mail: sergey.samsonov@biotec.tu-dresden.de.

Manuscript received 21 Apr. 2016; revised 6 Sept. 2016; accepted 22 Sept. 2016. Date of publication 3 Oct. 2016; date of current version 2 Feb. 2018. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2016.2614821

matching SSEs on a small data set [24]. Approaches derived from graph theory like clique detection are often used for protein structure comparison and recognition [25], [26], [27], [28], [29], [30], [31], [32], in computational biochemistry and genomics analysis for proteins [33] as well as for small molecules, e.g., [34]. By screening the whole PDB with our database-driven infrastructure, we found that IL-8 and PI-3 possess significant structural similarities. Comparison of their structures, key interactive residues, and electrostatic properties led to some new structure-based understanding of the unusual pepsin resistance of IL-8.

In order to obtain experimental insights into pepsin/IL-8 interaction in solution, an *in vitro* pepsin activity assay employing a short peptide substrate was used for detecting any inhibitory effect of IL-8 binding to the enzyme, and mass spectrometry (MS) based method was employed to detect any IL-8 degradation by the enzyme under standard and denaturizing solvent conditions. While no inhibition of enzyme activity was witnessed in the activity assay, enzymatic degradation of IL-8 was obtained only when exposed to mild denaturizing conditions. Furthermore, ^1H - ^{15}N HSQC (Heteronuclear Single Quantum Coherence) NMR experiments were carried out to characterize the residues participating in pepsin/IL-8 interactions. To summarize, in this work, we present computational and experimental results which propose the putative molecular mechanisms for the intrinsic IL-8 resistance to pepsin.

2 METHODS

2.1 Computational Methods

Our database-driven infrastructure contains PDBML [35] documents from the PDB [36], a unique worldwide repository for 3D data of biological macromolecules, which recently surpassed 100,000 entries. The basic design, pre-processing workflow and set-up is published elsewhere [37]. We extended our PDBML document processing workflow to enhance our database-driven infrastructure by storing SSE data as described below. In July 2015, we stored all available 110,919 PDBML documents at that time using IBMs DB2 Enterprise Server Edition v10.5 for Linux, Unix and Windows (LUW).

To determine SSEs, the processing workflow was extended by considering solely PDBML documents with protein structures. Only from these documents, the 3D atomic coordinates were extracted to apply DSSP [38], a widely accepted standard for secondary structure assignment [39]. Note that only SSEs consisting of four or more residues were considered. Inside the database, each SSE gets an automatically generated record ID (SSE-ID) assigned and is represented - depending on the number of residues the SSE consists of - at least as one penta-peptide (PP). For example, a SSE with nine residues ($n \dots n+8$) would be broken down ('fragmented') into two PPs $n \dots n+4$ and $n+4 \dots n+8$ with one overlapping residue $n+4$. A pseudo-algorithm with the benefits of PPs is presented in Supplementary Fig. A, see Appendix, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2016.2614821>. A database search for spatial, connectivity-independent arrangement of SSEs in the 3D space, is not directly based on matching SSEs but their underlying PPs. To query all available protein structures for matching SSEs efficiently, all PPs

Fig. 1. Simplified processing workflow to find matching SSEs in the 3D space. Here, four user-given SSEs (denoted with A, B, C, and D) were taken for example and their corresponding pairwise PPs (A'B', A'C', etc.) determined. Those PPs were looked up in the database. Any such PP-tuple can be schematically denoted as orange tubes. The white asterisk (**) at the source of the coordinate system represents the center of mass of the α -carbons belonging to both PPs. The dots in each tube represent the centered α -carbon, whereas one (of K') lies on the x-axis and the other (of L') on the xy-plane (denoted by the black dotted line). The search algorithm returns all such PP-tuples (K'L') that are located similarly in the hash table like A'B', A'C', etc., which is denoted as temporary match or PP-pair. Such set of PP-pairs represents the input for clique detection to determine whether they are unique and complete. If so, a protein structure is found having similarly arranged SSEs in the 3D space like the input SSEs (A, B, C, D). Both matching structures are last superimposed fully computationally (see text below for details).

were pre-processed according the reference frame generation of geometric hashing [20].

The basic idea of the reference frame generation here is to rearrange PPs in space and in addition to determine all its geometric invariants, see Fig. 1. The reference frame is a hash table, enabling fast SSE- (i.e., PP-) comparisons at the recognition stage [23]. Here, a record in the hash table basically contains any two ordered (K'L') PPs (denoted as PP-tuple) assigned with the SSE-ID it belongs to. Before inserting such a PP-tuple into the hash table, rigid transformations (translations, rotations) were applied on it, so that the center of mass of all ten α -carbons of K' and L' lies on the origin of the 3D coordinate system, the central α -carbon of K' on the x-axis and the central α -carbon of L' on the xy-plane. With that assignment of K'L' to the coordinate system, we generated our connectivity-independent 'building blocks' for the PP-search in our database. That assignment is done for each permutation of K'L' (i.e., K'L', L'K'), so that at runtime (i.e., recognition stage of geometric hashing) no special treatment of SSE/PP arrangement is required anymore.

For the recognition stage of geometric hashing, a C++-program directly accessing the DB2 database through an application programming interface (API) was developed. For the database search by efficiently scanning the hash table for matches, these input-SSEs were grouped into sets of two SSE, and each time their PP-tuples were looked up and fired against the hash table (A'B', A'C', A'D', B'C', ..., C'D'), see Fig. 1, upper part.

A temporary match was found when the root mean square deviation (rmsd) of the PP-tuple (based on the five α -carbons for each PP) in the table fall below 10 Å. The C++-binary also directly accesses the API of the CLIQUER

package [40] for determining whether found subsets of PPs in the database match not only all the input PPs, but also that this match is unique, see Fig. 1, lower part. This means, any found PP matches one and only one input PP (PP-pair). Such PP-pairs were then superimposed - using the API of BLAS (Basic Linear Algebra Subprograms) e.g., for rigid transformations. A final match was found only when the rmsd of the superimposed PPs did not exceed 3.0 Å for the α -carbons. Both overlapping PDB structures were finally stored in a PDB file.

For the calculations and visualization of the surface-mapped electrostatic potential of both PI-3 and IL-8, BIOVIA (formerly Accelrys) Discovery Studio was used.

2.2 Experimental Materials and Methods

2.2.1 Pepsin Activity Assays

The inhibition of pepsin activity by IL-8 was tested *in vitro* with the recombinantly produced human proteases pepsin A (Uniprot P00790; amino acids 16-388) and pepsin C (Uniprot P20142; amino acids 15-388), respectively.

The fluorescence lifetime-based protease activity assays were conducted in 384-well plates (Black Microtiter 384 Plate, round well, catalog number 95040020; Thermo Electron Oy, Vantaa, Finland) in a final volume of 25 μ l.

For the pepsin A activity assay, a buffer comprising 50 mM sodium acetate at pH 3.4, 33 mM sodium chloride, 1 mM ethylenediaminetetraacetic acid (EDTA), 1 mM dithiothreitol (DTT) and 0.05 percent (w/v) 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS) was used. For the pepsin C activity assay, the buffer comprised 50 mM glycine/HCl at pH 2.5, 33 mM sodium chloride, 1 mM ethylenediaminetetraacetic acid (EDTA), 1 mM dithiothreitol (DTT) and 0.05 percent (w/v) 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS).

The fluorescently labeled peptide Ac-Arg-Gly-Leu-Leu-Val-Trp-Gly-Leu-Ser-Gly-Cys(PT14)-Lys-NH₂ was purchased from Biosyntan (Berlin, Germany). Within the substrate sequence given above, PT14 is the acridone fluorescence lifetime probe. IL-8 was added to the plate containing either 0.3 nM pepsin A or 0.5 nM pepsin C, respectively, and incubated for one hour at room temperature. The change in the fluorescence lifetime of PT14 was used for monitoring pepsin activity.

The enzymatic reaction for evaluating the degree of inhibition of the enzyme was started by adding 1 μ M of the substrate and the fluorescence lifetime of the samples were measured after one hour incubation at room temperature.

Fluorescence lifetime measurements based on time-correlated single-photon counting (TCSPC) were conducted on an Ultra Evolution (TECAN, Maennedorf, Switzerland) fluorescence lifetime reader. The excitation light source was a semiconductor laser at 405 nm, producing picosecond light pulses with a repetition frequency of 10 MHz. The emission was collected through a 450-nm band-pass filter with 25 nm bandwidth. The measurement time per well was 1 s, yielding approximately 1000 counts in the peak channel. The parameters used as assay readout were the fluorescence lifetimes obtained by fitting the recorded fluorescence intensity decays with a monoexponential curve model done automatically by the control software of the instruments [41].

A Novartis proprietary inhibitor of pepsin A and C activity was used to check for the performance of the assay. For this inhibitor IC₅₀ values of 0.7 and 4.3 μ M in the pepsin A and C assays, respectively, were determined in-house previously.

2.2.2 LC/MS-Based Investigations of IL-8 Stability in the Presence of Pepsin

IL-8 at 40 μ M was incubated with either pepsin A or C at 1 nM in Eppendorf tubes at room temperature overnight. The respective assay buffers are reported in the section above.

The reducing agent TCEP (tris(2-carboxyethyl)phosphine) was added following the overnight incubation to a final concentration of 50 μ M.

After the incubation under reducing conditions at room temperature for 90 minutes, samples were analyzed by MS on an Acquity UPLC system (Waters Corporation, USA). The instrument was equipped with an Acquity UPLC BEH300 C18 (2.1 x 50 mm 1.7 μ m) column. The samples were run with a 5 to 98 percent (v/v) Eluent B gradient in 4.4 minutes and a flow rate of 0.8 ml/min. Eluent A was water with 0.05 percent trifluoroacetic acid (TFA) whereas Eluent B was acetonitrile with 0.04 percent TFA.

IL-8 and cleavage products were detected by the UV absorption recorded at 214 nm and their mass determined by ESI Q-TOF MS. For the sequence determination based on the MS data the "Protein Analysis Work Sheet" software (PAWS, Proteometrics Inc., New York, USA) was used.

Pepsin activity assays as well as IL-8 stability investigations in the presence of the pepsin were carried out under conditions previously justified and used in other experimental studies of this system [3], [42], [43].

2.2.3 NMR Spectroscopy-¹H-¹⁵N-HSQC Chemical Shift Perturbation Assay

Human pepsin was commercially not available, so porcine pepsin was taken (Sigma Aldrich: P7012-1G, Lot Number: SLBJ4999V). Note that the structure of porcine pepsin closely resembles the structure of human pepsin [44] with 84 percent sequence identity and 0.6 Å rmsd of superimposed structures [45]. The pepsin powder was directly dissolved in the measuring buffer (20 mM NaH₂PO₄, 50 mM NaCl, pH 7.0). The final protein concentration was 142 μ M as determined by UV/VIS spectroscopy ($E_{280\text{nm}} = 51300 \text{ M cm}^{-1}$) [46] using a NanoDrop device (Thermo Scientific, NanoDrop products, Wilmington, DE, USA).

The 77 amino acid variant of uniformly ¹⁵N-labelled human IL-8 (hIL-8) was produced by heterologous over expression in *E. coli* cells using a minimal medium and purified following the protocol published in [47]. 250 μ l of the hIL-8 preparation were dialyzed overnight against 1 l of the measuring buffer. The protein concentration was controlled by UV/VIS spectroscopy ($E_{280\text{nm}} = 7240 \text{ M cm}^{-1}$) [46] using a NanoDrop device (Thermo Scientific, NanoDrop products, Wilmington, DE, USA) and further adjusted to be 142 μ M.

¹H-¹⁵N-HSQC spectra were recorded for different molar ratios of hIL-8 monomer to porcine pepsin (1:0, 100:1, 50:1, 20:1, 10:1, 5:1 and 2:1) on a Bruker Avance III 700 MHz spectrometer equipped with a cryoprobe using the standard Bruker pulse sequence fhscqcf3gpph with 192 increments in

TABLE 1
Brief Summary of All the Determined SSEs by DSSP
with Absolute and Relative Occurrences

SSE type (DSSP annotation)	Absolute occurrence	Relative occurrence (in %)
π -helix (I)	2,807	0.05
3 ₁₀ -helix (G)	740,236	13.79
α -helix (H)	1,938,462	36.11
β -strand (E)	2,686,686	50.05
Sum	5,368,191	100.00

the indirect dimension (acquisition time: 138 ms), 16 scans per increment and a relaxation delay of 1 s. For the last two titration steps, only 128 increments in the indirect dimension were recorded and the number of scans was increased to 80 and 1696, respectively. The signals were assigned based on the published data in [46]. For each titration step, the weighted chemical shift changes $\Delta\delta$ for every assigned signal in the $^1\text{H}/^{15}\text{N}$ -HSQC of hIL-8 were calculated according to [48].

$$\Delta\delta = \sqrt{(\delta_{H,0} - \delta_{H,i})^2 + \left(\frac{\delta_{N,i} - \delta_{N,0}}{5}\right)^2},$$

where $\delta_{N,i}$ represents the ^{15}N - chemical shift and $\delta_{H,i}$ the ^1H chemical shift measured at the i th titration step.

3 RESULTS AND DISCUSSIONS

3.1 *In Silico* Analysis of IL-8 and PI-3 Structures

From the 110,619 available PDB structures as of July 2015, 107,793 of them contain one or more polypeptide chain(s) of any length. DSSP assigned to 106,348 PDB structures at least one SSE and 5,368,191 in total, see Table 1.

Our hash table contains 654,142,535 'fragmented' PPs. A run using four (three β -strands, α -helix) SSEs of IL-8 (PDB ID: 1IL8) resulted in 3,154 PDB output files with superimposed PPs/SSEs, distributed to 2,383 PDB structures. Also, due to the SSE fragmentation, some structures have up to ten (PDB IDs: 4XIX, 3LK4) or even 16 (PDB ID: 3IA0) SSE matchings. One of the results that caught our attention was for a complex containing pepsin and a pepsin inhibitor termed PI-3 (PDB ID: 1F34).

For the *in silico* analysis of the structures with the PDB ID 1IL8 and 1F34 we used the original residue numbering conventions from Uniprot [49] and denote corresponding PDB numbering for clarity, see Supplementary Fig. B in the Appendix, available online. For example, the PDB numbering for 1IL8 would start at Uniprot position 28 (denoted in capital letters for each molecule).

PI-3 was found to be among the top-scored hits. In particular, it was ranked as 48 in total and as 9 among the hits not belonging to the IL-8-like fold (Supplementary Table A, Appendix, available online). There are a couple of essential key motifs of PI-3 being important for the inhibition. They are listed in Table 2 with spatially corresponding IL-8 counterparts (if available), being analyzed more detailed below.

The superposition of the IL-8 and PI-3 structures shows high similarity. The α -carbons rmsd values of superimposed PPs are 2.9 Å and 2.2 Å for automatically - i.e., fully

TABLE 2
Important Structural Fragments for the Pepsin Inhibition
of PI-3 and Their IL-8 Counter-Parts (if Available) Using
Uniprot Sequence Numbering

PI-3	IL-8 (automatically superimposed)	IL-8 (manually superimposed)
β -strand 21-23 (QFL)	-	-
β -strand 24-28 (FSMST)	N-term. β -strand 51-55 (ELRVI)	C-term. β -strand 74-78 (RELCL)
130 (K)	92 (K)	81 (K)
poly-PRO 161I-163I (PPP)	46 (P)	80 (P)

computationally calculated through our C++ -application using BLAS (Basic Linear Algebra Subprograms) to determine the transformation matrix. The matching with the best rmsd value for PI-3 and IL-8 is shown in Fig. 2a.

The 3D structure analysis of the complex between porcine pepsin and PI-3 has revealed the key residues of PI-3 participating in aspartic protease inhibition [8]. This means, the binding site consisting of residues 24-28 - FSMST - of the N-terminal strand of PI-3 make antiparallel β -sheet hydrogen bonding interactions with residues 129-133 - ELSIT - of

Fig. 2. Superimposed structures of PI-3 (red) and IL-8 (blue). Residues 51-55 of the N-terminal β -strand of IL-8 are colored in yellow and 28-42 are removed, both for clarity. The β -strand (res. 24-28) of PI-3 also being the binding site to pepsin is denoted in green. Superimposed structures on panel (a) are created computationally according to Fig. 1 and on panel (b) manually, reflecting HSQC NMR results (see text below). In the latter model, the C-terminal β -strand (res. 74-78) of IL-8 covers the β -strand (res. 24-28) of PI-3, which corresponds better with the weighted chemical shift difference in Fig. 6.

Fig. 3. The electrostatic potentials of IL-8 (panel a) and PI-3 (panel b) mapped on the solvent exclusion surface. Each arrow denotes the location of pepsin looking at the binding site of IL-8/PI-3.

pepsin [8]. This segment of pepsin is part of a β -hairpin that forms part of the S1'-S3' binding sites in pepsin [50]. For the fully computationally obtained superimposition of IL-8 and PI-3, the β -strand encompassing residues 24-28 of PI-3 match the N-terminal β -strand of IL-8 with the corresponding residues ELRVI at sequence positions 51-55 with α -carbons rmsd ranging from 1.2 Å to 2.7 Å, see Fig. 2 (panel A).

Next, we compared the electrostatic potential of the superimposed PI-3 and IL-8 (computational model), particularly both regions pointing to the pepsin binding site. Fig. 3 shows the electrostatic potential mapped on the solvent exclusion surface for 1IL8 (left side) and 1F34 (right side).

Comparing the calculated surface-mapped electrostatic potential for both molecules reveals that they share similar electrostatic properties suggesting that IL-8 can form a putative interface with the pepsin similar to the one formed by PI-3.

There are few residues that are essential for the pepsin inhibition of PI-3. Besides the poly-proline helix (159-163), also the N-terminal residues 21-23 (QFL) were observed to be important for the inhibition of various aspartic proteases [51]. These residues occupy enzymatic subsites P1', P2' and P3' of the pepsin and so block the access of potential substrates to the S1' to S3' binding pockets in the enzyme [8]. That glutamine of PI-3 is located near the catalytic aspartic acids on positions 91 and 274 of pepsin, whereas the α -amino group of glutamine is reported to be nearest to the carboxylate oxygen atom of residue 274 that forms a hydrogen bond with a catalytically important water molecule in the active enzyme [5], [8]. The closest distance of an IL-8 residue (computational model) to the carboxylate oxygen atom of residue 274 is about 11.9 Å. The superimposition of IL-8 and PI-3 structures reveals no counterpart in IL-8 corresponding to the tri-peptide QFL of PI-3, neither for the computational nor the manual model (see below) suggesting that there is no direct putative interference of IL-8 with the pepsin active site.

Inhibitor residues 161, 162 and 163 of the C-terminal poly-proline helix (159-163) are crucial for optimal binding to porcine pepsin, whereas residue 159 is not significant for the inhibition [51]. The interactions between the 349's α -hairpin loop of pepsin and this poly-proline helix of PI-3 may also contribute to the specificity of PI-3 [8]. Apparently, there are no corresponding IL-8 poly-proline segments, but

Fig. 4. Dose-response curves for IL-8 and a known reference inhibitor obtained with *in vitro* activity assay for human pepsin A (panel a) and C (panel b).

there is a viral chemokine structure (PDB ID: 1ZXT) containing such a tripeptide of prolines in the structurally corresponding region. However, the role of proline 160 in PI-3 could be potentially performed by proline 46 (computational model) in the 3_{10} -helix of IL-8, whereas the closest distance between the two pyrrolidines of P160 and P46 is less than 2.9 Å. Moreover, another important residue of PI-3, K130, has a counterpart residue (computational model), K92, in IL-8 (NZ-NZ-distance is less than 4.0 Å).

Using four SSEs of IL-8 as input for a database search of connectivity-independent SSEs obtained two structural matches regarding a pepsin-PI-3-complex. Both matches differ structurally solely by slightly shifted SSEs. From the literature, we know the key residues of pepsin and PI-3 which are responsible for the inhibition. In most cases, we identified matching counterpart regions/residues in IL-8, which may partially mimic the function of PI-3. Additionally, we calculated the electrostatic potentials of IL-8 and PI-3. We then compared both electrostatic potentials with IL-8 superimposed on PI-3 - particularly along the binding region of PI-3 to pepsin - and observed similar electrostatic properties.

3.2 Inhibition of Pepsin Activity by IL-8 Using Pepsin Assays

The inhibition of human pepsin A and pepsin C activity by IL-8 was tested *in vitro* in a dose-dependent manner using a protease activity assay. Fig. 4 shows the dose-response curves (DRCs) for a known reference inhibitor (red lines) and for IL-8 (black lines).

The DRCs for the reference inhibitors result in IC_{50} values of 0.7 and 4.3 μ M for pepsin A and C, respectively, prove that the assays performed well. For IL-8, a dose-dependency of the pepsin activities on the IL-8 concentrations was not obtained up to a concentration of 40 μ M. Hence, the IL-8 is not inhibiting the two enzymes under the

TABLE 3
Primary Structures of Peptides Detected and Identified
by LC/MS after IL-8 Incubation with Pepsin C

Peptide	Primary structure (one letter code)	Residues	Detected mass (Da)
IL-8	avlp ^r SAKELRCQCIKTYSKPFHPKFIK ELRVIESGPHCANTEIIVKLS*DGREL *CLDPKENW*VQ*RVVEKFLKRAENS	23-99	8922
Prod. 1	avlp ^r SAKELRCQCIKTYSKPFHPKFIK ELRVIESGPHCANTEIIVKLSDGREL	23-76	6151
Prod. 2	avlp ^r SAKELRCQCIKTYSKPFHPKFIK ELRVIESGPHCANTEIIVKLSDGREL LDPKENW	23-84	7137
Prod. 3	DGRELCLDPKENWVQ	72-86	1802

reported assay conditions, i.e., when a short peptide is used for monitoring enzymatic activity *in vitro*. The stability of IL-8 in the presence of pepsin A and C under reducing and non-reducing conditions was tested by incubation the chemokine with either of the proteases and the subsequent analysis of the reaction mixtures by LC/MS.

The analysis revealed that under the chosen reducing assay conditions corresponding to the native state of IL-8, IL-8 is not cleaved by either of the proteases. In contrast, under the same assay conditions and in the presence of the reducing agent TCEP, pepsin C but not pepsin A was able to cleave the chemokine, suggesting that IL-8 could be cleaved in the denaturated state. These results are in agreement with previously obtained data for IL-8 resistance to pepsin [3].

The detected IL-8 cleavage is a slow process; it was witnessed only after incubating the chemokine with the protease overnight at room temperature. In Table 3, the detected cleavage products and the resulting cleavage sites are shown. The asterisks (*) indicate the cleavage sites in IL-8 by pepsin C found in this study.

3.3 IL-8/Pepsin Interactions Analysis by ¹H-¹⁵N HSQC NMR Spectroscopy

Fig. 5 shows the overlaid ¹H-¹⁵N HSQC NMR spectra at different molar ratios of pepsin to hIL-8 monomer. Clearly, pepsin concentration dependent chemical shift changes of hIL-8 signals are observed, e.g., for the backbone amide signals of S71 and L76. Fig. 6 shows the resulting weighted chemical shift differences plotted along the sequence of hIL-8.

Caused by the relatively low concentration of the pepsin protein stock solution, only a twofold molar excess of pepsin over hIL-8 monomer was feasible for this titration study. It should be noted that this was not enough to fully drive the chemical shift perturbations to saturation. The amide signals of N83/Y40 and K94/R87 could not be unambiguously followed during titration and were not included in Fig. 6 but at least one signal of each pair showed a shift. For a similar reason, I66/I37 and K30/C34 were also not included, but regarding the first pair both signals showed a shift, whereas regarding the second pair no shift was observed. L32, A29 and Y40 could not be detected at all. In summary, the pronounced chemical shift perturbations, which clearly depend on pepsin concentration, suggest an interaction of hIL-8 and porcine pepsin in solution. Due to the high structural similarity between pepsins A and C,

Fig. 5. Overlaid ¹H-¹⁵N-HSQC NMR spectra at different porcine pepsin to hIL-8 (77 residues) monomer molar ratios (0:1, 1:10, 1:5, 1:2). Clearly, pepsin concentration dependent chemical shift changes of hIL-8 signals are observed, e.g., for S71 and L76.

these results also suggest a similar interface for IL-8/pepsin C interactions.

The computationally obtained superimposition depicted in Fig. 2a displays the overlap between residues 51-55 of the N-terminal β -strand of IL-8 with residues 24-28 of the β -strand of PI-3, which is also the binding site of the pepsin. Compared to Fig. 6, we do not observe NMR signals around positions 51-55 which may indicate an interaction of IL-8 with the pepsin. Thus, based on the structural superimposition depicted in Fig. 2a we used BIOVIA Discovery Studio in order to adjust the automatically obtained superimposition by manually translating and rotating IL-8 molecule. The obtained superimposition (Fig. 2b) is different from the initial one and also contains a good structural match of SSEs of PI-3 and IL-8. In contrast to the superimposition in Fig. 2a, residues 74-78 of the C-terminal β -strand of IL-8 cover the β -strand (res. 24-28) of PI-3 and hence overlap with the pepsin binding site of PI-3. This superimposition yields a better understanding of the pepsin resistance of IL-8 at molecular level than the automatic one and suggests the pepsin binding region that is in a good agreement with the one inferred from the NMR experiments.

4 CONCLUSION

In this work, we aimed at understanding the structural basis of IL-8 resistance to pepsin. *In silico*, we combined geometric hashing and clique detection to search the entire PDB for spatially matching SSEs. We demonstrated that both approaches can be successfully applied for identifying SSEs in the 3D space taking the entire PDB into account. IL-8 was the queried protein with the four SSE being screen through the database. One of the most interesting database matches was PI-3 of the nematode *A. suum*. Since IL-8 resists pepsin cleavage, we first analyzed the structural similarities of IL-8

Fig. 6. Weighted chemical shift difference $\Delta\delta$ (in ppm) for IL-8 at different titration points (molar ratios of porcine pepsin to hIL-8) plotted along the amino acid sequence of hIL-8 (77 residues).

and PI-3. It turned out that some essential PI-3 key players for inhibition can be partially ‘mimicked’ by IL-8. Second, we decided to cross-check our promising *in silico* observations experimentally using pepsin assays and HSQC NMR experiments. Indeed, in the former experiment, IL-8 did not inhibit pepsin, but resisted cleavage. In the later experiment, we clearly proved strong binding of pepsin and IL-8 with the binding interface partially in agreement with the one we proposed computationally. Our molecular model and experimental findings provide structural insights into the IL-8 resistance complementing the experimental data previously obtained by other approaches [3].

ACKNOWLEDGMENTS

The study was supported in part by the Deutsche Forschungsgemeinschaft (TRR-SFB 67, project A6).

REFERENCES

- [1] M. Baggiolini, A. Walz, and S. L. Kunkel, “Neutrophil-activating peptide-1/interleukin 8, a novel cytokine that activates neutrophils,” *J. Clin. Investigation*, vol. 84, no. 4, pp. 1045–1049, Oct. 1989.
- [2] W. Schnitzel, U. Mondschein, and J. Besemer, “Monomer-dimer equilibria of interleukin-8 and neutrophil-activating peptide 2. Evidence for IL-8 binding as a dimer and oligomer to IL-8 receptor B,” *J. Leukocyte Biol.*, vol. 55, no. 6, pp. 763–770, Jun. 1994.
- [3] A. Maheshwari, W. Lu, W. C. Guida, R. D. Christensen, and D. A. Calhoun, “IL-8/CXC ligand 8 survives neonatal gastric digestion as a result of intrinsic aspartyl proteinase resistance,” *Pediatric Res.*, vol. 57, no. 3, pp. 438–444, Mar. 2005.
- [4] K. Girdwood and C. Berry, “The disulphide bond arrangement in the major pepsin inhibitor PI-3 of *Ascaris suum*,” *FEBS Lett.*, vol. 474, no. 2/3, pp. 253–255, Jun. 2000.
- [5] D. R. Davies, “The structure and function of the aspartic proteinases,” *Annu. Rev. Biophys. Biophys. Chem.*, vol. 19, pp. 189–215, Jun. 1990.
- [6] J. S. Fruton, “A history of pepsin and related enzymes,” *Quarterly Rev. Biol.*, vol. 77, no. 2, pp. 127–147, Jun. 2002.
- [7] B. M. Dunn, “Structure and mechanism of the pepsin-like family of aspartic peptidases,” *Chem. Rev.*, vol. 102, no. 12, pp. 4431–4458, Dec. 2002.
- [8] K. K. S. Ng, et al., “Structural basis for the inhibition of porcine pepsin by *Ascaris* pepsin inhibitor-3,” *Nat. Struct. Biol.*, vol. 7, no. 8, pp. 653–657, Aug. 2000.
- [9] J. Willenbücher, W. Höfle, and R. Lucius, “The filarial antigens Av33/Ov33-3 show striking similarities to the major pepsin inhibitor from *Ascaris suum*,” *Mol. Biochem. Parasitology*, vol. 57, no. 2, pp. 349–352, Feb. 1993.
- [10] M. Li, et al., “The aspartic proteinase from *Saccharomyces cerevisiae* folds its own inhibitor into a helix,” *Nat. Struct. Biol.*, vol. 7, no. 2, pp. 113–117, Feb. 2000.
- [11] G. M. Abu-Erreish and R. J. Peanasky, “Pepsin inhibitors from *Ascaris lumbricoides*. Isolation, purification, and some properties,” *J. Biol. Chem.*, vol. 249, no. 5, pp. 1558–1565, Mar. 1974.
- [12] T. Kageyama, “Molecular cloning, expression and characterization of an *Ascaris* inhibitor for pepsin and cathepsin E,” *Eur. J. Biochem.*, vol. 253, no. 3, pp. 804–809, May 1998.
- [13] N. S. A. Krushna, C. Shiny, G. Manokaran, S. Elango, and R. B. Narayanan, “Immune responses to recombinant *Brugia malayi* pepsin inhibitor homolog (Bm-33) in patients with human lymphatic filariasis,” *Parasitology Res.*, vol. 108, no. 2, pp. 407–415, Feb. 2011.
- [14] M. J. Valler, J. Kay, T. Aoyagi, and B. M. Dunn, “The interaction of aspartic proteinases with naturally-occurring inhibitors from actinomycetes and *Ascaris lumbricoides*,” *J. Enzyme Inhibition*, vol. 1, no. 1, pp. 77–82, 1985.
- [15] F. D. Finkelman, et al., “Cytokine regulation of host defense against parasitic gastrointestinal nematodes: Lessons from studies with rodent models,” *Annu. Rev. Immunology*, vol. 15, pp. 505–533, Apr. 1997.
- [16] W. C. Gause, J. F. Urban Jr., and M. J. Stadecker, “The immune response to parasitic helminths: Insights from murine models,” *Trends Immunology*, vol. 24, no. 5, pp. 269–277, May 2003.
- [17] C. Bazzocchi, C. Genchi, S. Paltrinieri, C. Lecchi, M. Mortarino, and C. Bandi, “Immunological role of the endosymbionts of *dirofilaria immitis*: The wolbachia surface protein activates canine neutrophils with production of IL-8,” *Veterinary Parasitology*, vol. 117, no. 1/2, pp. 73–83, Nov. 2003.
- [18] L. Proudfoot, “Parasitic helminths tip the balance: Potential anti-inflammatory therapies,” *Immunology*, vol. 113, no. 4, pp. 438–440, Dec. 2004.
- [19] F. H. Falcone, A. G. Rossi, R. Sharkey, A. P. Brown, D. I. Pritchard, and R. M. Maizels, “*Ascaris suum*-derived products induce human neutrophil activation via a G protein-coupled receptor that interacts with the interleukin-8 receptor pathway,” *Infection Immunity*, vol. 69, no. 6, pp. 4007–4018, Jun. 2001.

- [20] H. J. Wolfson and I. Rigoutsos, "Geometric hashing: An overview," *IEEE Trans. Comput. Sci. Eng.*, vol. 4, no. 4, pp. 10–21, Apr. 1997. [Online]. Available: http://www.cse.unr.edu/~mircea/Teaching/cs485_685/Readings/Wolfson97.pdf
- [21] X. Pennec and N. Ayache, "An $O(n^2)$ algorithm for 3D substructure matching of proteins," in *Proc. 1st Int. Workshop Shape Pattern Matching Comput. Biol.*, 1994, pp. 25–40.
- [22] A. C. Wallace, N. Borkakoti, and J. M. Thornton, "TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites," *Protein Sci.*, vol. 6, no. 11, pp. 2308–2323, Nov. 1997.
- [23] X. Pennec and N. Ayache, "A geometric algorithm to find small but highly similar 3D substructures in proteins," *Bioinf.*, vol. 14, no. 6, pp. 516–522, 1998.
- [24] L. Holm and C. Sander, "3-D lookup: Fast protein structure database searches at 90% reliability," in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 1995, vol. 3, pp. 179–187.
- [25] E. M. Mitchell, P. J. Artymiuk, D. W. Rice, and P. Willett, "Use of techniques derived from graph theory to compare secondary structure motifs in proteins," *J. Mol. Biol.*, vol. 212, no. 1, pp. 151–166, Mar. 1990.
- [26] H. M. Grindley, P. J. Artymiuk, D. W. Rice, and P. Willett, "Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm," *J. Mol. Biol.*, vol. 229, no. 3, pp. 707–721, Feb. 1993.
- [27] K. Mizuguchi and N. Go, "Comparison of spatial arrangements of secondary structural elements in proteins," *Protein Eng.*, vol. 8, no. 4, pp. 353–362, Apr. 1995.
- [28] I. Koch, T. Lengauer, and E. Wanke, "An algorithm for finding maximal common subtopologies in a set of protein structures," *J. Comput. Biol.*, vol. 3, no. 2, pp. 289–306, Summer 1996.
- [29] E. J. Gardiner, P. J. Artymiuk, and P. Willett, "Clique-detection algorithms for matching three-dimensional molecular structures," *J. Mol. Graph. Model.*, vol. 15, no. 4, pp. 245–253, Aug. 1997.
- [30] A. Harrison, F. Pearl, R. Mott, J. Thornton, and C. Orengo, "Quantifying the similarities within fold space," *J. Mol. Biol.*, vol. 323, no. 5, pp. 909–926, Nov. 2002.
- [31] N. Weskamp, D. Kuhn, E. Hüllermeier, and G. Klebe, "Efficient similarity search in protein structure database by k-clique hashing," *Bioinf.*, vol. 20, no. 10, pp. 1522–1526, Jul. 2004.
- [32] B. Kolbeck, P. May, T. Schmidt-Goenner, T. Steinke, and E.-W. Knapp, "Connectivity independent protein-structure alignment: A hierarchical approach," *BMC Bioinf.*, vol. 7, Nov. 2006, Art. no. 510.
- [33] S. Butenko and W. E. Wilhelm, "Clique-detection models in computational biochemistry and genomics," *Eur. J. Oper. Res.*, vol. 173, no. 1, pp. 1–17, Aug. 2006.
- [34] P. Willett, "Searching for pharmacophoric patterns in databases of three-dimensional chemical structures," *J. Mol. Recognit.*, vol. 8, no. 5, pp. 290–303, Sep./Oct. 1995.
- [35] J. D. Westbrook, N. Ito, H. Nakamura, K. Henrick, and H. M. Berman, "PDBML: The representation of archival macromolecular structure data in XML," *Bioinf.*, vol. 21, no. 7, pp. 988–992, Apr. 2005.
- [36] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley, "The worldwide protein data bank (wwPDB): Ensuring a single, uniform archive of the PDB data," *Nucleic Acids Res.*, vol. 35, pp. D301–D303, Jan. 2007.
- [37] G. Anders and M. Nicola, "Managing the protein data bank with DB2 pureXML," *developerWorks*, 2011. [Online]. Available: <http://www.ibm.com/developerworks/data/library/techarticle/dm-1109proteinatadb2purexml/>
- [38] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.
- [39] J. Martin, G. Letellier, A. Marin, J.-F. Taly, A. G. de Breverin, and J.-F. Gibrat, "Protein secondary structure assignment revisited: A detailed analysis of different assignment methods," *BMC Struct. Biol.*, vol. 5, Sep. 2005, Art. no. 17.
- [40] S. Niskanen and P. R. J. Østergård, "Cliques user's guide, version 1.0," Commun. Laboratory, Helsinki Univ. Technol., Espoo, Finland, Tech. Rep. T48, 2003.
- [41] K. Doering, G. Meder, M. Hinnenberger, J. Woelcke, L. M. Mayr, and U. Hassiepen, "A fluorescence lifetime-based assay for protease inhibitor profiling on kallikrein 7," *J. Biomol. Screen.*, vol. 14, no. 1, pp. 1–9, Jan. 2009.
- [42] T. Kageyama, "Pepsinogens, progastricsins, and prochymosins: Structure, function, evolution, and development," *Cell Mol. Life Sci.*, vol. 59, no. 2, pp. 288–306, Feb. 2002.
- [43] A. Maheshwari, et al., "Interleukin-8/CXCL8 forms an autocrine loop in fetal intestinal mucosa," *Pediatric Res.*, vol. 56, no. 2, pp. 240–249, Aug. 2004.
- [44] A. R. Sielecki, A. A. Fedorov, A. Boodhoo, N. S. Andreeva, and M. N. G. James, "Molecular and crystal structures of monoclinic porcine pepsin refined at 1.8 Å resolution," *J. Mol. Biol.*, vol. 214, no. 1, pp. 143–170, Jul. 1990.
- [45] M. Fujinaga, M. M. Chernai, N. I. Tarasova, S. C. Mosimann, and M. N. G. James, "Crystal structure of human pepsin and its complex with pepstatin," *Protein Sci.*, vol. 4, no. 5, pp. 960–972, May 1995.
- [46] E. Gasteiger, et al., "Protein identification and analysis tools on the ExPASy server," *The Proteomics Protocols Handbook*, J. M. Walker, Ed., Humana Press, New York, pp. 571–607, 2005.
- [47] A. Pichert, et al., "Characterization of the interaction of interleukin-8 with glycosaminoglycans by solution NMR spectroscopy and molecular modelling," *Glycobiology*, vol. 22, no. 1, pp. 134–145, Jan. 2012.
- [48] E. S. Seo, et al., "Interaction of human β -defensin 2 (HBD2) with glycosaminoglycans," *Biochemistry*, vol. 49, no. 49, pp. 10486–10495, Dec. 2010.
- [49] Consortium UniProt, "UniProt: A hub for protein information," *Nucleic Acids Res.*, vol. 43, pp. D204–D12, Jan. 2015.
- [50] M. Fujinaga, M. M. Cherney, N. I. Tarasova, P. A. Bartlett, J. E. Hanson, and M. N. G. James, "Structural study of the complex between human pepsin and a phosphorus-containing peptidic transition-state analog," *Acta Crystallographic D Biol. Crystallographic*, vol. 56, no. Pt. 3, pp. 272–279, Mar. 2000.
- [51] R.E. Moose, et al., "Analysis of binding interactions of pepsin inhibitor-3 to mammalian and malarial aspartic proteases," *Biochemistry*, vol. 46, no. 49, pp. 14198–14205, Dec. 2007.

Gerd Anders received the diploma degree in informatics from Humboldt Universität zu Berlin, in 2005. At present, he is a consultant at ARS Computer und Consulting GmbH (Munich). His research interests include structural bioinformatics, computational geometry, efficient data management, and processing as well as cognitive computing.

Ulrich Hassiepen received the PhD degree in biochemistry from the University of Technology (RWTH) Aachen, Germany, where he also studied chemistry. In 1998, he joined Novartis Pharma AG in Basel, Switzerland, where he is working in preclinical research at the Center of Proteomic Chemistry of the Novartis Institutes for BioMedical Research.

Stephan Theisgen received the diploma degree in biochemistry from Martin-Luther-University Halle/Wittenberg, Germany, in 2008 and the PhD degree from the University of Leipzig, Germany, in 2012. After two years as postdoc in the research group of Prof. Huster at the University of Leipzig, he joined ZoBio B.V. in Leiden (The Netherlands), where he is working in structure-driven drug discovery.

Stephan Heymann received the diploma in physics from the University of Yerevan and the PhD degree in molecular biology in Berlin. He then led academic and industry-oriented projects in Germany and Russia. Later on, the research focus was in the fields of recombinant techniques, immunology, and bioinformatics. Currently, his main interests concern the catalytic impact of natural minerals on metabolism.

Daniel Huster received the diploma and PhD degrees from the University of Leipzig before he spent a 2-year postdoctoral stay with Iowa State University. He returned to Leipzig and Halle, where he led independent junior research groups before he was called to become full professor of medical biophysics with the University of Leipzig. His research focusses on membrane proteins, protein-glycosaminoglycan interaction, amyloids, and biological tissue.

Lionel Muller received the bachelor's degree in biochemistry from the University of Strasbourg, ULP, France, where he also studied biochemistry. In 1996, he joined Novartis Pharma AG in Basel, Switzerland, where he is working in pre-clinical research at the Center of Proteomics Chemistry of the Novartis Institutes for BioMedical Research.

Sergey A. Samsonov received the MS degree from Saint-Petersburg State Polytechnical University, in 2006 and the PhD degree from the Dresden University of Technology, in 2009. He is currently a postdoctoral researcher in the Group of Structural Bioinformatics at Dresden University of Technology. His studies mainly focus on molecular dynamics-based approaches applied to protein-protein, protein-saccharide, and protein-DNA interactions.

Tania Panigada received the MSc degree in medical biology from the University of Lausanne (UNIL), Switzerland, where she also studied biology. In 2011, she joined Novartis Pharma AG in Basel, Switzerland, where she is working in pre-clinical research at the Analytical Science & Imaging platform of the Novartis Institutes for BioMedical Research.

▷ For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.

Teil III

Zusammenfassung

Dissertation zur Erlangung des akademischen Grades

Dr. rer. med.

Informatik-gestützte Fallstudie zur Analyse von Proteinstrukturen am Beispiel der Pepsinresistenz von Interleukin-8

eingereicht von
Gerd Anders

angefertigt am
Institut für Medizinische Physik und Biophysik
der Universität Leipzig

betreut von
Prof. Dr. rer. nat. Daniel Huster

Monat und Jahr (der Einreichung)
März 2018

Diese Publikationspromotion implementiert und bewertet hauptsächlich algorithmische Verfahren für die strukturelle Bioinformatik im Kontext von Konzeptstudien (proof of concepts). Bei der Kombination dieser Verfahren steht insbesondere deren (struktur-)biologische Eignung zur Analyse von Protein(struktur)en im Vordergrund. Folglich werden Komplexitäten (d. h. Laufzeiteigenschaften) und Speicherbedarf der Verfahren weniger ausführlich charakterisiert.

Die Datengrundlage bilden Proteinstrukturen, welche weltweit in Laboren experimentell ermittelt werden. Die daraus resultierenden Ergebnisse werden bis auf wenige Ausnahmen in einer einzelnen Datei in einem vorgegebenen Format gespeichert und dem Protein Data Bank (PDB) Konsortium übergeben. Aufgrund des historisch bedingten (Lochkarten-)Formats und der Vielzahl (100000+) an Dateien sind effiziente, vielseitige Suchen in diesem gesamten Datenbestand nicht bzw. kaum durchführbar. Einige Lösungsansätze mit dem Ziel von schnelleren Suchen wurden damals in der wissenschaftlichen Literatur publiziert. Wirklich durchsetzen konnten sich diese Ansätze in der Praxis nicht, vor allem weil sie technisch aufwendig ausgearbeitet waren und ein tiefgreifendes Verständnis ihrer internen Datenstrukturen erforderten.

Im Rahmen dieser Publikationspromotion musste nach Möglichkeit zunächst ein hocheffizientes Verfahren zur Aufbereitung und Speicherung von Daten aus PDB Dateien entwickelt und ggf. evaluiert werden. Die zugrunde liegende wissenschaftliche Fragestellung lautete:

(1) Kann man die Gesamtheit an Proteinstrukturen der PDB mit Mitteln der Informatik derart speichern, sodass diese allumfassend und effizient abgefragt werden können?

Als Ergebnis von (1) wurde eine neuartige, Datenbank-gestützte Infrastruktur für umfassende und vielseitige strukturelle Analysen von Proteinen geschaffen. Diese Infrastruktur ist eine entscheidende Grundlage für darauf aufbauende algorithmische Verfahren zum Vergleich von Proteinstrukturen.

Im Wesentlichen handelte es sich um teilweise bekannte algorithmische Verfahren wie geometrisches Hashing³ und Clique Detection⁴. Bekannte, für den Praxiseinsatz nachteilige Eigenschaften dieser Verfahren sind deren algorithmische Komplexität (d. h. Laufzeiteigenschaften) oder der nicht unerhebliche Speicherbedarf. Diese Nachteile wurden u. a. in der wissenschaftlichen Literatur thematisiert, mit der Erkenntnis, dass die Verfahren im Allgemeinen kaum bzw. nur sehr beschränkt anwendbar sind. Im Rahmen dieser Publikationspromotion mussten die Nachteile folglich erst durch geeignete Maßnahmen abgeschwächt werden. Die daraus resultierende wissenschaftliche Fragestellung lautete:

(2) Wie kann man diese Verfahren auf atomarer Ebene auf die Gesamtheit der Proteinstrukturen anwendbar machen (wenn überhaupt)?

Als Ergebnis von (2) konnten geeignete Maßnahmen identifiziert und implementiert werden, so dass man damit grundsätzlich sämtliche Proteine strukturell vergleichen konnte. Ob dieses Vergleichsverfahren jedoch für die Analyse von Proteinen aus strukturellbiologischer Sicht pragmatisch war, blieb im Sinne der o. g. Konzeptstudie zunächst offen. Folglich wurden die von diesem Verfahren erzielten Ergebnisse in Form von überlagerten Proteinstrukturen stichprobenartig untersucht, um eine nicht abschließende Einschätzung bzw. Bewertung vorzunehmen. Damit galt es die folgenden wissenschaftlichen Fragestellungen zu beantworten:

(3) Eignen sich die stichprobenartigen Ergebnisse, d. h. die überlagerten Proteinstrukturen, aus dem Vergleichsverfahren um qualitativ hochwertige strukturellbiologische Analysen durchzuführen? Lassen sich damit ggf. auch offene strukturellbiologische Fragestellungen beantworten?

Anhand eines konkreten Anwendungsbeispiels von zwei überlagerten Proteinstrukturen konnten zahlreiche strukturellbiologische Besonderheiten identifiziert werden. Bei den beiden Proteinen handelte es sich um Interleukin-8 als Testprotein (PDB ID: 1IL8) und einen Proteinkomplex. Die Festlegung von Interleukin-8 erfolgte aus subjektiven Erwägungen, da Interleukin-8 bereits in vorherigen Projekten involviert war. Der Proteinkomplex (PDB ID: 1F34) war das Ergebnis einer umfassenden (Datenbank-)Suche durch das Vergleichsverfahren und bestand aus Pepsin und einem Pepsin Inhibitor-3.

In der wissenschaftlichen Literatur gab es bereits mehrere Beobachtungen hinsichtlich der Pepsinresistenz von Interleukin-8. Darüber hinaus auch Bestrebungen die ungewöhnliche Pepsinresistenz von Interleukin-8 auf molekularer Ebene zu verstehen. Aus strukturellbiologischer Sicht ergaben die wenigen dazu in der wissenschaftlichen Literatur beschriebenen Versuche keine plausiblen Erklärungsansätze für die Pepsinresistenz. Das hier implementierte und verwendete Vergleichsverfahren brachte dagegen zahlreiche strukturelle Anhaltspunkte, die zusammenfassend eine schlüssige Hypothese ergaben. Anhand dieser Hypothese wurden dann zielgerichtet Laborexperimente durchgeführt, um die Hypothese zu evaluieren. Im Zusammenspiel der in-silico Strukturanalyse mit den aufschlußreichen Laborergebnissen konnte die Pepsinresistenz von Interleukin-8 auf molekularer Ebene plausibel erklärt werden.

³Unter geometrischen Hashing kann man sich eine geeignete Aufteilung einer großen Datenmenge von Punktkoordinaten in vielen kleine Datenmengen vorstellen, um insbesondere die Suchen nach Koordinaten zu beschleunigen.

⁴Ein Graph hat eine Clique, wenn zu einer Anzahl von Ecken (Vertices) jede Ecke mit einer anderen über eine Kante verbunden ist. Ein Dreieck ist bspw. eine Clique, ein Rechteck nicht.

Zu (1): Vorteile einer Datenbank-gestützten PDB-Infrastruktur

Seit mehr als 45 Jahren werden die Daten zu aufgedeckten Proteinstrukturen vom PDB-Konsortium der Allgemeinheit zur freien Verfügung gestellt. Die damalige Verwendung von Lochkarten hat dabei das Schema vorgegeben. D. h., die Gesamtheit an Daten von einer Proteinstruktur wird in einem durch Lochkarten bedingten, semi-strukturiertem Format abgespeichert. Da viele Anwendungen aus historischen Gründen darauf aufbauen, haben Schemaänderungen unter Umständen weitreichende Auswirkungen. Dieses Schema ist für einen umfassenden Zugriff auf PDB-Daten aus heutiger Sicht jedoch nur eingeschränkt verwendbar, z. B. für Data Mining.

Mit der Einführung eines XML-basierten Schemas - genannt PDBML (PDB Markup Language) - durch das PDB Konsortium wurden diese Einschränkungen vollständig behoben. Die zumindest damalige Popularität von XML (Extensible Markup Language) spiegelte sich auch in der Unterstützung durch diverse Datenbanksysteme, allen voran DB2 von IBM, wider. Insbesondere durch zweckmäßige Transformation der PDBML-Dokumente konnten diese Daten vollständig in DB2 gespeichert werden, sodass sämtliche PDB Daten integriert und allumfassend (durch SQL und XQuery) zugreifbar wurden. Damit wurden einerseits vielseitige Abfragen, bspw. zu Analyse Zwecken, möglich. Andererseits konnte durch zusätzliche Integration von Daten aus anderen Quellen diese mit PDB Daten einfach verknüpft werden (mittels SQL und XQuery), sodass weit umfangreichere Analysen ermöglicht wurden. Diese pragmatische Datenintegration (nicht nur) von PDB Daten in DB2 wird hier im weitesten Sinne als PDB2 (PDB und DB2) Infrastruktur bezeichnet.

Zu (2): Eigenschaften der neuen Vergleichsmethode für Proteinstrukturen

In der wissenschaftlichen Literatur gibt es zahlreiche Verfahren, Proteinstrukturen vergleichen. Viele dieser Algorithmen wenden eine Datenreduktion an. Sekundärstrukturelemente werden bspw. gerne als Vektor repräsentiert, da dieser nur zwei 3D-Koordinaten benötigt. Diese Vereinfachung ermöglicht zwar sehr schnelle Vergleiche und einen reduzierten Ressourcenbedarf, jedoch werden die Ergebnisse, bspw. in Form von überlagerten Proteinstrukturen, durch die Vereinfachung zumeist ungenauer.

Im vorliegenden Fall wurden Sekundärstrukturelemente als Grundlage für den Vergleich von Proteinstrukturen verwendet. Anschließend wurden bestimmte Atomkoordinaten ($C\alpha$ -Atome) der Sekundärstrukturelemente entsprechend extrahiert und für geometrisches Hashing aufbereitet. Geometrisches Hashing ist ein robustes Verfahren um bspw. Atome mit geringem Abstand zueinander im 3D-Raum effizient zu ermitteln. Der effiziente Datenzugriff auf die entsprechenden Atomkoordinaten wird dabei durch das Datenbanksystem unterstützt, obwohl geometrisches Hashing einen signifikanten Mehrbedarf an Speicherplatz hat. Insgesamt wurde auf Basis von geometrischem Hashing ein Datenbank-gestütztes Verfahren implementiert, welches sämtliche Proteinstrukturen - bzw. in diesem Fall deren Sekundärstrukturelemente - auf atomarer Ebene vergleichen kann. Durch den Vergleich werden Proteinstrukturen identifiziert, deren Sekundärstrukturelemente sich im 3D-Raum überlagern oder nur einen geringen Abstand zueinander aufweisen. Wie noch gezeigt wird, können solche Überlagerungen aus biologischer Sicht bedeutsam sein und zu neuen Erkenntnissen auf struktureller und u. U. auch auf funktioneller Ebene führen.

Zu (3): Stichprobenartige Evaluation der Ergebnisse aus der neuen Vergleichsmethode

Aus der Sicht der Informatik konnte bei (2) aufgezeigt werden, dass durch geschickte Datenaufbereitung und die Verwendung eines Datenbanksystems das rechenintensive und Ressourcen-beanspruchende geometrische Hashing grundsätzlich auf sämtliche Proteinstrukturen der PDB angewendet werden kann, und zwar ohne die Notwendigkeit einer Datenreduktion. Aus (struktur-)biologischer Sicht stellt sich zunächst die Frage, ob sich die in (2) ermittelten Ergebnisse in Form von überlagerten Proteinstrukturen biologisch weiter auswerten lassen.

Zu Evaluationszwecken wurden diverse Testläufe mit Interleukin-8 durchgeführt. Vier (drei β -

Stränge und die C-terminale α -Helix) der fünf Sekundärstrukturelemente von Interleukin-8 (PDB ID: 1IL8) dienten dabei als Eingabe. Trotz stringenter Suchparameter (wie rmsd⁵-Werte der überlagerten Proteinstrukturen) wurden meist tausende Überlagerungen im gesamten PDB Datenbestand gefunden. Besonders vielversprechende Ergebnisse (hinsichtlich rmsd-Werte, PDB-Beschreibungstext, etc.) wurden dann stichprobenartig analysiert.

Die strukturelle Bewertung der überlagerten Sekundärstrukturelemente der Stichproben erfolgte visuell unter Verwendung eines Molekülbetrachters. Eine dieser Stichproben enthielt eine Überlagerung von Interleukin-8 mit einem Proteinkomplex bestehend aus Pepsin und einem Pepsin Inhibitor-3 (PDB ID: 1F34). Ungewöhnlich waren hier insbesondere die verschiedenen Topologien der überlagerten Sekundärstrukturelemente von Interleukin-8 und dem Pepsin Inhibitor-3. Durch entsprechende Literaturrecherchen ergaben sich auf molekularer Ebene zahlreiche und plausible Hinweise, warum Interleukin-8 von Pepsin nicht gespalten wird.

Aufgrund dieser strukturellen Hinweise konnten zielgerichtete Laborversuche durchgeführt werden, um die Hinweise zu verifizieren. Insgesamt kann man festhalten, dass die Laborergebnisse grundsätzlich die Erkenntnisse aus der in-silico Analyse bestätigten. Damit konnte zumindest eine bislang ungelöste Frage zur Pepsinresistenz von Interleukin-8 auf molekularbiologischer Ebene beantwortet werden. In den letzten Jahren gab es zwar bereits Bestrebungen diese Frage zu beantworten. Diese Versuche trugen nur unwesentlich zu neuen Erkenntnissen bei, weil sie aufgrund fehlender (struktureller) Anhaltspunkte nicht zielgerichtet durchgeführt wurden.

Insgesamt wurde der Nachweis erbracht, dass einerseits geometrisches Hashing ohne Datenreduzierung auf den gesamten, aufbereiteten PDB Datenbestand angewendet werden kann. Andererseits wurden in einem konkreten Fall durch geometrisches Hashing überlagerte Proteinstrukturen gefunden, welche die entscheidende Grundlage für die Beantwortung einer offenen strukturellen Frage waren. Neben dem o. g. Proteinkomplex gab es weitere interessante Überlagerungen von Proteinstrukturen, welche jedoch zeit- und aufwandsbedingt (bisher noch) nicht weiter analysiert wurden.

⁵root-mean-square-deviation (deutsch: Wurzel aus der mittleren quadratischen Abweichung)

Teil IV

Anhang

Literaturverzeichnis

- [Abagyan88] Abagyan, R. A.; Maiorov, V. N.: A Simple Qualitative Representation of Polypeptide Chain Folds: Comparison of Protein Tertiary Structures.
J. Biomol. Struct. Dynam., 5, 1267-1279, (1988)
- [Abagyan89] Abagyan, R. A.; Maiorov, V. N.: An Automatic Search for Similar Spatial Arrangements of α -Helices and β -Strands in Globular Proteins.
J. Biomol. Struct. Dynam., 6, 1045-1279, (1989)
- [Abu-Erreish74] Abu-Erreish, G. M.; Peanasky, R. J.: Pepsin Inhibitors from *Ascaris lumbricoides*.
J.Biol. Chem., 249, 1558-1565, (1974)
- [Abyzov07] Abyzov, A.; Ilyin, V. A.: A comprehensive analysis of non-sequential alignments between all protein structures.
BMC Struct. Biol., 7, 78-97, (2007)
- [Achard01] Achard, F.; Vaysseix, G.; Barillot, E.: XML, bioinformatics and data integration.
Bioinformatics, 17, 115-125, (2001)
- [Adler06] Adler, S.; Cochrane, R.; Morar, J. F.; Spector, A.: Technical context and cultural consequences of XML.
IBM Systems Journal, 45, 207-223, (2006)
- [Albeck06] Albeck, S.; Alzari, P.; Andreini, C.; Banci, L.; Berry, I. M.; Bertini, I.; Cambillau, C.; Canard, B.; Carter, L.; Cohen, S. X.; Diprose, J. M.; Dym, O.; Esnouf, R. M.; Felder, C.; Ferron, F.; Guillemot, F.; Hamer, R.; Ben Jelloul, M.; Laskowski, R. A.; Laurent, T.; Longhi, S.; Lopez, R.; Luchinat, C.; Malet, H.; Mochel, T.; Morris, R. J.; Moulinier, L.; Oinn, T.; Pajon, A.; Peleg, Y.; Perrakis, A.; Poch, O.; Prilusky, J.; Rachedi, A.; Ripp, R.; Rosato, A.; Silman, I.; Stuart, D. I.; Sussman, J. L.; Thierry, J. C.; Thompson, J. D.; Thornton, J. M.; Unger, T.; Vaughan, B.; Vranken, W.; Watson, J. D.; Whamond, G.; Henrick, K.: SPINE bioinformatics and data-management aspects of high-throughput structural biology.
Acta Cryst., D62, 1184-1195, (2006)
- [Alexander15] Alexander, S. P.; Davenport, A. P.; Kelly, E.; Marrion, N.; Peters, J. A.; Benson, H. E.; Faccenda, E.; Pawson, A. J.; Sharman, J. L.; Southan, C.; Davies, J. A.; CGTP Collaborators: The Concise Guide to PHARMACOLOGY 2015/16: G protein-coupled receptors.
Br. J. Pharmacol., 172, 5744-869, (2015)
- [Altman04] Altman, R. B.; Dugan, J. M.: Defining bioinformatics and structural bioinformatics.
Methods Biochem. Anal., 44, 3-14, (2004)
- [Alexandrov96a] Alexandrov, N. N.: SARFing the PDB.
Protein Eng., 9, 727-732, (1996)

- [Alexandrov96b] Alexandrov, N. N.; Fischer, D.: Analysis of Topological and Nontopological Structural Similarities in the PDB: New Examples With Old Structures. *Proteins*, 25, 354-365, (1996)
- [Anders11] Anders, G.; Nicola, M.: Managing the Protein Data Bank with DB2 pureXML. *Managing the Protein Data Bank with DB2 pureXML*, (2011)
- [Anders16] Anders, G.; Hassiepen, U.; Theisgen, S.; Heymann, S.; Muller, L.; Panigada, T.; Huster, D.; Samsonov, S. A.: The Intrinsic Pepsin Resistance of Interleukin-8 Can be Explained from a Combined Bioinformatical and Experimental Approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, in press, (2016)
- [Anderson99] Anderson, E.; Bai, Z.; Bischof, C.; Blackford, S.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; Sorensen, D.: *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, (1999)
- [Anfinsen77] Anfinsen, C. B.: Principles that govern the folding of protein chains. *Science*, 181, 223-230, (1977)
- [Ashburner00] Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, J. M.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G.: Gene Ontology: tool for the unification of biology. *Nature Genet.*, 25, 25-29, (2000)
- [Aung07] Aung, Z.; Tan, K.-L.: Rapid retrieval of protein structures from databases. *Drug Discovery Today*, 12, 732-739, (2007)
- [B06] : Steering the future of computing. *Nature*, 440, 383, (2006)
- [Bachelier14] Bachelier, F.; Ben-Baruch, A.; Burkhardt, A. M.; Combadiere, C.; Farber, J. M.; Graham, G. J.; Horuk, R.; Sparre-Ulrich, A. H.; Locati, M.; Luster, A. D.; Mantovani, A.; Matsushima, K.; Murphy, P. M.; Nibbs, R.; Nomiyama, H.; Power, C. A.; Proudfoot, A. E.; Rosenkilde, M. M.; Rot, A.; Sozzani, S.; Thelen, M.; Yoshie, O.; Zlotnik, A.: International Union of Basic and Clinical Pharmacology. [corrected]. LXXXIX. Update on the extended family of chemokine receptors and introducing a new nomenclature for atypical chemokine receptors.. *Pharmacol. Rev.*, 66, 1-79, (2014)
- [Bacon02] Bacon, K.; Baggiolini, M.; Broxmeyer, H.; Horuk, R.; Lindley, I.; Mantovani, A.; Matsushima, K.; Murphy, P.; Nomiyama, H.; Oppenheim, J.; Rot, A.; Schall, T.; Tsang, M.; Thorpe, R.; Van Damme, J.; Wadhwa, M.; Yoshie, O.; Zlotnik, A.; Zoon, K.: Chemokine/chemokine receptor nomenclature. *J. Interferon Cytokine Res*, 22, 1067-1068, (2002)
- [Baggiolini89] Baggiolini, M.; Walz, A.; Kunkel, S. L.: Neutrophil-activating Peptide-1/Interleukin 8, a Novel Cytokine That Activates Neutrophils. *J. Clin. Invest.*, 84, 1045-1049, (1989)
- [Baggiolini92] Baggiolini, M.; Clark-Lewis, I.: Interleukin-8, a chemotactic and inflammatory cytokine. *FEBS*, 307, 97-101, (1992)

- [Baggiolini94] Baggiolini, M.; Loetscher, P.; Moser, B.: INTERLEUKIN-8 AND THE CHEMOKINE FAMILY.
Int. J. Immunopharmac., 17, 103-108, (1994)
- [Baggiolini97] Baggiolini, M.; Dewald B.; Moser, B.: Human chemokines: an update.
Annu. Rev. Immunol., 15, 675-705, (1997)
- [Bairoch00] Bairoch, A.; Apweiler, R.: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.
Nucleic Acids Res., 28, 45-48, (2000)
- [Baldwin91] Baldwin, E. T.; Weber, I. T.; Charles, R. S.; Xuan, J.-C.; Apella, E.; Yamada, M.; Matsushima, K.; Edwards, B. F. P.; Clore, G. M.; Gronenborn, A. M.; Wlodawer, A.: Crystal structure of interleukin 8: Symbiosis of NMR and crystallography.
Proc. Natl. Acad. Sci., 88, 502-506, (1991)
- [Banaszak00] Banaszak, L. J.: Foundations of Structural Biology.
Elsevier, (2003)
- [Baxevanis03] Baxevanis, A. D.: The Molecular Biology Database Collection: 2003 Update.
Nucleic Acids Res., 31, 1-12, (2003)
- [Baxevanis06] Baxevanis, A. D.: The Importance of Biological Databases in Biological Discovery.
Curr. Prot. Bioinformatics, Chapter 1: Unit1.1, 1-5, (2006)
- [Bayat02] Bayat, A.: Science, medicine, and the future: Bioinformatics.
BMJ, 324, 1018-1022, (2002)
- [Bazan97] Bazan, J. F.; Bacon, K. B.; Hardiman, G.; Wang, W.; Soo, K.; Rossi, D.; Greaves, D. R.; Zlotnik, A.; Schall, T. J.: A new class of membrane-bound chemokine with a CX_3X motif.
Nature, 385, 640-644, (1997)
- [Bazzocchi03] Bazzocchi, C.; Genchi, C.; Paltrinieri, S.; Lecchi, C.; Mortarino, M.; Bandi, C.: Immunological role of the endosymbionts of *Dirofilaria immitis*: the *Wolbachia* surface protein activates canine neutrophils with production of IL-8.
Vet. Parasitol., 117, 73-83, (2003)
- [Bergholz01] Bergholz, A.; Heymann, S.; Schenk, J. A.; Freytag, J. C.: Biological sequences Integrated: a relational database approach.
Acta Biotheor., 49, 145-159, (2001)
- [Berkamp17] Berkamp, S.; Park, S. H.; De Angelis, A. A.; Marassi, F. M.; Opella, S. J.: Structure of monomeric Interleukin-8 and its interactions with the N-terminal Binding Site-I of CXCR1 by solution NMR spectroscopy.
J. Biomol. NMR, 69, 111-21, (2017)
- [Berman92] Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeney, T.; Hsieh, S.-H.; Srinivasan, A. R.; Schneider, B.: The Nucleic Acid Database: A comprehensive relational database of three-dimensional structures of nucleic acids.
Biophys. J., 63, 751-759, (1992)
- [Berman00a] Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J.: The Protein Data Bank and the challenge of structural genomics.
Nat. Struct. Biol., 7, 957-959, (2000)

- [Berman00b] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E.: The Protein Data Bank. *Nucleic Acids Res.*, 28, 235-42, (2000)
- [Berman07] Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L.: The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of the PDB data. *Nucleic Acids Res.*, 35, D301-D303, (2007)
- [Berman08] Berman, H. M.: The Protein Data Bank: a historical perspective. *Acta Cryst.*, A64, 88-95, (2008)
- [Berman09] Berman, H. M.; Westbrook, J. D.; Gabanyi, M. J.; Tao, W.; Shah, R.; Kouranov, A.; Schwede, T.; Arnold, K.; Kiefer, F.; Bordoli, L.; Kopp, J.; Podvinec, M.; Adams, P. D.; Carter, L. G.; Minor, W.; Nair, R.; La Baer, J.: The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Res.*, 37, D365-D368, (2009)
- [Berman12] The Protein Data Bank at 40: reflecting on the past to prepare for the future: Berman, H. M.; Kleywegt, G. J.; Nakamura, H.; Markley, J. L.. *Structure*, 20(3), 391-6, (2012)
- [Berman14] Berman, H. M.; Kleywegt, G. J.; Nakamura, H.; Markley, J. L.: The Protein Data Bank archive as an open data resource. *J. Comput. Aided Mol. Des.*, 28(10), 1009-14, (2014)
- [Bernstein77] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer E. F. jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M.: The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.*, 112, 535-542, (1977)
- [Bertone01] Bertone, P.; Gerstein, M.: Integrative Data Mining: The New Direction in Bioinformatics. *Eng. Med. Biol. Mag.*, 20, 33-40, (2001)
- [Bhambhri05] Bhambhri, A.: Firing up the hybrid engine. *IBM database magazin*, 10, <http://www.ibmdatabasemag.com/ShowArticle.jhtml?articleID=167100937>, (2005)
- [Bhat01] Bhat, T. N.; Bourne, P.; Feng, Z.; Gilliland, G.; Jain, S.; Ravichandran, V.; Schneider, B.; Schneider, K.; Thanki, N.; Weissig, H.; Westbrook, J.; Berman, H. M.: The PDB data uniformity project. *Nucleic Acids Res.*, 29, 214-218, (2001)
- [Bhattacharyya04] Bhattacharyya, R.; Pal, D.; Chakrabarti, P.: Disulfide bonds, their stereospecific environment and conservation in protein structure. *Protein Eng. Des. Sel.*, 29, 795-808, (2004)
- [Billeter08] Billeter, M.; Wagner, G.; Wüthrich, K.: Solution NMR structure determination of proteins revisited. *J. Biomol. NMR*, 42(3), 155-8, (2008)
- [Bjorkman87] Bjorkman, P. J.; Saper, M. A.; Samraoui, B.; Bennett, W. S.; Strominger, J. L.; Wiley, D. C.: Structure of the human class I histocompatibility antigen, HLA-A2. *Nature*, 329, 506-512, (1987)

- [Bolton70] Bolton, W.; Perutz, M. F.: Three dimensional fourier synthesis of horse deoxyhaemoglobin at 2.8 Å resolution.
Nature, 228, 551-2, (1970)
- [Bonecchi09] Bonecchi, R.; Galliera, E.; Borroni, E. M.; Corsi, M. M.; Locati, M.; Mantovani, A.: Chemokines and chemokine receptors: an overview.
Front. Biosci., 14, 540-551, (2009)
- [Bourne03] Bourne, P. E.; Address, K. J.; Bluhm, W. F.; Chen, L.; Deshpande, N.; Feng, Z.; Fleri, W.; Green, R.; Merino-Ott, J. C.; Townsend-Merino, W.; Weissig, H.; Westbrook, J.; Berman, H. M.: The distribution and query systems of the RCSB Protein Data Bank.
Nucleic Acids Res, 32, D223-D225, (2003)
- [Bry03] Bry, F.; Kröger, P.: A Computational Biology Database Digest: Data, Data Analysis, and Data Management.
Distributed and Parallel Databases, 13, 7-42, (2003)
- [Bukhman01] Bukhman, Y. V.; Skolnick, J.: BioMolQuest: integrated database-based retrieval of protein structural and functional information.
Bioinformatics, 17, 468-478, (2001)
- [Burley13] Burley, S. K.: PDB40: The Protein Data Bank celebrates its 40th birthday.
Biopolymers, 99(3), (2013)
- [Burley17a] Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S.: Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive.
Methods Mol. Biol., 1607, 627-641, (2017)
- [Burley17b] Burley, S. K.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S.: Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive.
Methods Mol. Biol., 1607, 627-41, (2017)
- [Burley18] Burley, S. K.; Berman, H. M.; Christie, C.; Duarte, J. M.; Feng, Z.; Westbrook, J.; Young, J. Zardecki, C.: RCSB Protein Data Bank: Sustaining a living digital resource that enables breakthroughs in scientific research and biomedical education.
Protein Sci., 27, 316-330, (2018)
- [Butenko06] Butenko, S.; Wilhelm, W. E.: Clique-detection models in computational biochemistry and genomics.
Eur. J. Oper. Res., 173, 1-17, (2006)
- [Butler99] Butler, D.: IBM promises scientists 500-fold leap in supercomputing power...
Nature, 402, 705-706, (1999)
- [Buttler02] Buttler, D.; Coleman, M.; Critchlow, T.; Fileto, R.; Han, W.; Liu, L.; Pu, C.; Rocco, D.; Xiong, L.: Quering Multiple Bioinformatics Information Sources: Can Semantic Web Research Help?.
ACM SIGMOD, 31, 59-64, (2002)
- [C97] : Structure and the genome.
Nat. Struct. Biol., 4, 329-330, (1997)

- [Callaway96] Callaway, J.; Cummings, M.; Deroski, B.; Esposito, P.; Forman, A.; Langdon, P.; Libeson, M.; McCarthy, J.; Sikora, J.; Xue, D.; Abola, E.; Bernstein, F.; Manning, N.; Shea, R.; Stampf, D.; Sussman, J.: Protein Data Bank Contents Guide: Atomic coordinate entry format description.
Brookhaven National Laboratory, (1996)
- [Carugo06] Carugo, O.: Rapid Methods for Comparing Protein Structures and Scanning Structure Databases.
Curr. Bioinform., 1, 75-83, (2006)
- [Carugo07] Carugo, O.: Recent Progress in Measuring Structural Similarity Between Proteins.
Curr. Protein Pept. Sci., 8, 219-241, (2007)
- [Chandonia06] Chandonia, J.-M.; Brenner, S. E.: The Impact of Structural Genomics: Expectations and Outcomes.
Science, 311, 347-351, (2006)
- [Chang94] Chang, W.; Shindyalov, I. N.; Pu, C.; Bourne, P. E.: Design and application of PDBlib, a C++ macromolecular class library.
Comput. Appl. Biosci., 10, 575-586, (1994)
- [Chen03] Chen, J. Y.; Carlis, J. V.: Similiar_Join: Extending DBMS with a Bio-specific Operator.
Proc. of the 2003 ACM Symposium on Applied Computing, 109-114, (2003)
- [Chruszcz08] Chruszcz, M.; Wlodawer, A.; Minor, W.: Determination of Protein Structures - A Series of Fortunate Events.
Biophys. J., 95, 1-9, (2008)
- [Clark-Lewis94] Clark-Lewis, I.; Dewald, B.; Loetscher, M.; Moser, B.; Baggiolini, M.: Structural Requirements for Interleukin-8 Function Identified by Design of Analogs and CXC Chemokine Hybrids.
J. Biol. Chem., 269, 16075-16081, (1994)
- [Clark-Lewis95] Clark-Lewis, I.; Kim, K.-S.; Rajarathman, K.; Gong, J.-H.; Dewald, B.; Moser, B.; Baggiolini, M.; Sykes, B. D.: Structure-activity relationships of chemokines.
J. Leukoc. Biol., 57, 703-711, (1995)
- [Clore95] Clore, G. M.; Gronenborn, A. M.: Three-dimensional structures of α and β chemokines.
FASEB J., 9, 57-62, (1995)
- [Cochrane10] Cochrane, G. R.; Galperin, M. Y.: The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources.
Nucleic Acids Res., 38, D1-4, (2010)
- [Davidson95] Davidson, S. B.; Overton, C.; Buneman, P.: Challenges in Integrating Biological Data Sources.
J. Comput. Biol., 2, 557-572, (1995)
- [Davies90] Davies, D. R.: THE STRUCTURE AND FUNCTION OF THE ASPARTIC PROTEINASES.
Annu. Rev. Biophys. Biophys. Chem., 19, 189-215, (1990)

- [Dawson17] Dawson, N. L.; Lewis, T. E.; Das, S.; Lees, J. G.; Lee, D.; Ashford, P.; Orengo, C. A.; Sillitoe, I.: CATH: an expanded resource to predict protein function through structure and sequence.
Nucleic Acids Res., 45(D1), D289-D295, (2017)
- [Demir17] Demir, Ö.; Jeong, P. U.; Amaro, R. E.: Full-length p53 tetramer bound to DNA and its quaternary dynamics.
Oncogene, 36(10), 1451-1460, (2017)
- [Deshpande05] Deshpande, N.; Address, K. J.; Bluhm, W. F.; Merino-Ott, J. C.; Townsend-Merino, W.; Zhang, Q.; Knezevich, C.; Xie, L.; Feng, Z.; Green, R. K.; Flippen-Anderson, J. L.; Westbrook, J.; Berman, H. M.; Bourne, P. E.: The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema.
Nucleic Acids Res., 33, D233-D237, (2005)
- [Dodge98] Dodge, C.; Schneider, R.; Sander, C.: The HSSP database of protein structure-sequence alignments and family profiles.
Nucleic Acids Res., 26, 313-315, (1998)
- [Dundas07] Dundas, J.; Binkowski, T. A.; DasGupta, B.; Liang, J.: Topology Independent Protein Structural Alignment.
BMC Bioinformatics, 8, 388-401, (2007)
- [Dunn02] Dunn, B. M.: Structure and Mechanism of the Pepsin-Like Family of Aspartic Peptidases.
Chem. Rev., 102, 4431-4458, (2002)
- [Dutta05] Dutta, S.; Berman, H. M.: Large Macromolecular Complexes in the Protein Data Bank: A Status Report.
Structure, 13, 381-388, (2005)
- [Dutta10] Dutta, S.; Zardecki, C.; Goodsell, D. S.; Berman, H. M.: Promoting a structural view of biology for varied audiences: an overview of RCSB PDB resources and experiences.
J. Appl. Cryst., 43, 1224-1229, (2010)
- [Efimov95] Efimov, A.V.: Structural Similarity between Two-layer α/β and β -proteins.
J. Mol. Biol., 245, 402-415, (1995)
- [Elmasri07] Elmasri, R.; Ji, F.; Fu, J.; Zhang, Y.; Raja, Z.: Modelling concepts and database implementation techniques for complex biological data.
Int. J. Bioinform. Res. Appl., 3, 366-388, (2007)
- [Fernandez02] Fernandez, E. J.; Lolis, E.: Structure, Function, and Inhibition of Chemokines.
Annu. Rev. Pharmacol. Toxicol., 42, 469-499, (2002)
- [FernandezSuarez13] Fernandez-Suarez, X. M.; Galperin, M. Y.: The 2013 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection.
Nucleic Acids Res., 41, D1-7, (2013)
- [FernandezSuarez14] Fernandez-Suarez, X. M.; Rigden, D. J.; Galperin, M. Y.: The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection.
Nucleic Acids Res., 41, D1-7, (2014)

- [Ferreira04] Ferreira, K. N.; Iverson, T. M.; Maghlaoui, K.; Barber, J.; Iwata, S.: Architecture of the photosynthetic oxygen-evolving center. *Science*, 303, 1831-1838, (2004)
- [Ferrari03] Ferrari, C.; Guerra, C.; Zanotti, G.: A grid-aware approach to protein structure comparison. *J. Parallel Distrib. Comput.*, 63, 728-737, (2003)
- [Finkelman97] Finkelman, F. D.; Shea-Donohue, T.; Goldhill, J.; Sullivan, C. A.; Morris, S. C.; Madden, K. B.; Gause, W. C.; Urban, J. F., Jr.: CYTOKINE REGULATION OF HOST DEFENSE AGAINST PARASITIC GASTROINTESTINAL NEMATODES: Lessons from Studies with Rodent Models. *Annu. Rev. Immunol.*, 15, 505-533, (1997)
- [Finn10] Finn, R. D.; Mistry, J.; Tate, J.; Coggill, P.; Heger, A.; Pollington, J. E.; Gavin, O. L.; Gunasekaran, P.; Ceric, G.; Forslund, K.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A.: The PFAM protein families database. *Nucleic Acids Res.*, 38, D211-D222, (2010)
- [Fliess02] Fliess, A.; Motro, B.; Unger, R.: Swaps in Protein Sequences. *Proteins*, 48, 377-387, (2002)
- [Fruton02] Fruton, J. S.: A History of Pepsin and Related Enzymes. *Q. Rev. Biol.*, 77, 127-147, (2002)
- [Gajda13] Gajda, M. J.: HPDB-Haskell library for processing atomic biomolecular structures in Protein Data Bank format. *BMC Res. Notes*, 6:483, (2013)
- [Gales13] Gales, D.; Clark, C.; Manne, U.; Samuel, T.: The Chemokine CXCL8 in Carcinogenesis and Drug Response. *ISRN Oncol*, 859154, (2013)
- [Galperin04] Galperin, M. Y.: The Molecular Biology Database Collection: 2004 Update. *Nucleic Acids Res.*, 32, D3-22, (2004)
- [Galperin05] Galperin, M. Y.: The Molecular Biology Database Collection: 2005 Update. *Nucleic Acids Res.*, 33, D5-24, (2005)
- [Galperin06] Galperin, M. Y.: The Molecular Biology Database Collection: 2006 Update. *Nucleic Acids Res.*, 34, D3-5, (2006)
- [Galperin07] Galperin, M. Y.: The Molecular Biology Database Collection: 2007 Update. *Nucleic Acids Res.*, 35, D3-4, (2007)
- [Galperin08] Galperin, M. Y.: The Molecular Biology Database Collection: 2008 Update. *Nucleic Acids Res.*, 36, D1-3, (2007)
- [Galperin09] Galperin, M. Y.; Cochrane, G. R.: Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Res.*, 37, D1-4, (2009)
- [Galperin11] Galperin, M. Y.; Cochrane, G. R.: The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.*, 39, D1-6, (2011)

- [Galperin12] Galperin, M. Y.; Fernandez-Suarez, X. M.: The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection.
Nucleic Acids Res., 40, D1-8, (2012)
- [Galperin15] Galperin, M. Y.; Rigden, D. J.; Fernandez-Suarez, X. M.: The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection.
Nucleic Acids Res., 43, D1-5, (2015)
- [Galperin17] Galperin, M. Y.; Fernandez-Suarez, X. M.; Rigden, D. J.: The 24th Nucleic Acids Research database issue: a look back and upcoming changes.
Nucleic Acids Res., 45, D1-11, (2017)
- [Gardiner97] Gardiner, E. J.; Artymiuk, P. J.; Willett, P.: Clique-detection algorithms for matching three-dimensional molecular structures.
J. Mol. Graph. Model., 15(4), 245-53, (1997)
- [Garin11] Garin, A.; Proudfoot, A. E. I.: Chemokines as targets for therapy.
Exp. Cell Res., 317(5), 602-12, (2011)
- [Gause03] Gause, W. C.; Urban, J. F., Jr.; Stadecker, M. J.: The immune response to parasitic helminths: insights from murine models.
Trends Immunol., 24, 269-277, (2003)
- [Gerstein99] Gerstein, M.: Building the future of biocomputing.
Nature, 399, 101, (1999)
- [Gerstein00] Gerstein, M.: Integrative database analysis in structural genomics.
Nat. Struct. Biol., 7, 960-963, (2000)
- [Gibrat96] Gibrat, J. F.; Madej, T.; Bryant, S. H.: Surprising similarities in structure comparison.
Curr. Opin. Struct. Biol., 6, 377-385, (1996)
- [Gilpin16] Gilpin, W.: PyPDB: a Python API for the Protein Data Bank.
Bioinformatics, 32(1), 159-60, (2016)
- [Girdwood00] Girdwood, K.; Berry, C.: The disulphide bond arrangement in the major pepsin inhibitor PI-3 of *Ascaris suum*.
FEBS, 23717, 252-256, (2000)
- [Gray90] Gray, P. M. D.; Paton, N. W.; Kemp, G. J. L.; Fothergill, J. E.: An object-oriented database for protein structure analysis.
Protein Eng., 3, 235-243, (1990)
- [Greer02] Greer, D. S.; Westbrook, J. D.; Bourne, P. E.: An ontology driven architecture for derived representations of macromolecular structure.
Bioinformatics, 18, 1280-1281, (2002)
- [Grindley93] Grindley, H. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P.: Identification of Tertiary Structure Resemblance in Proteins Using a Maximal Common Subgraph Isomorphism Algorithm.
J. Mol. Biol., 229, 707-721, (1993)
- [Hamelryck03] Hamelryck, T.; Manderick, B.: PDB file parser and structure class implemented in Python.
Bioinformatics, 19, 2308-2310, (2003)

- [Harrison02] Harrison, A.; Pearl, F.; Mott, R.; Thornton, J.; Orengo, C.: Quantifying the Similarities within Fold Space.
J. Mol. Biol., 323, 909-926, (2002)
- [Hawley94] Hawley, J. H.; Martzen, M. R.; Peanasky, R. J.: Proteinase Inhibitors in *Ascarida*.
Parasitol. Today, 10, 308-313, (1994)
- [Hawley95] Hawley, J. H.; Peanasky, R. J.: *Ascaris suum*: Are trypsin Inhibitors Involved in Species Specificity of *Ascarid* Nematodes.
Exp. Parasitol., 75, 112-118, (1995)
- [Heinemann95] Heinemann, U.; Hahn, M.: Circular permutations of protein sequence: not so rare?.
Trends Biochem. Sci., 20, 349-350, (1995)
- [Henrick08] Henrick, K.; Feng, Z.; Bluhm, W. F.; Dimitropoulos, D.; Doreleijers, J. F.; Dutta, S.; Flippen-Anderson, J. L.; Ionides, J.; Kamada, C.; Krissinel, E.; Lawson, C. L.; Markley, J. L.; Nakamura, H.; Newman, R.; Shimizu, Y.; Swaminathan, J.; Velankar, S.; Ory, J.; Ulrich, E. L.; Vranken, W.; Westbrook, J.; Yamashita, R.; Yang, H.; Young, J.; Yousufuddin, M.; Berman, H. M.: Remediation of the protein data bank archive.
Nucleic Acids Res., 36, D426-D433, (2008)
- [Hinneburg03] Hinneburg, A.; Lehner, W.: Database support for 3D-Protein Data Set Analysis.
Proc. of the 15th Int. Conf. on Scientific and Statistical Database Management (SSDBM'03), 161-170, (2003)
- [Hoffmann02] Hoffmann, E.; Dittrich-Breiholz, O.; Holtmann, H.; Kracht, M.: Multiple control of interleukin-8 gene expression.
J. Leukoc. Biol., 72, 847-855, (2002)
- [Holland08] Holland, R. C. G.; Down, T. A.; Pocock, M.; Prlic, A.; Huen, D.; James, K.; Foisy, S.; Dräger, A.; Yates, A.; Heuer, M.; Schreiber, M. J.: BioJava: an open-source framework for bioinformatics.
Bioinformatics, 24, 2096-2097, (2008)
- [Holm94] Holm, L.; Sander, C.: The FSSP database of structurally aligned protein fold families.
Nucleic Acids Res., 22, 3600-3609, (1994)
- [Holm95] Holm, L.; Sander, C.: 3-D Lookup: Fast Protein Structure Database Searches at 90% Reliability.
Proc. Int. Conf. Intell. Syst. Mol. Biol., 3, 179-187, (1995)
- [Holm96] Holm, L.; Sander, C.: The FSSP database: fold classification based on structure-structure alignment of proteins.
Nucleic Acids Res., 24, 206-209, (1996)
- [Holm97] Holm, L.; Sander, C.: Dali/FSSP classification of three-dimensional protein folds.
Nucleic Acids Res., 25, 231-234, (1997)
- [Howe08] Howe, D.; Rhee, S. Y.: The future of biocuration.
Nature (special issue: Big data), 455, 47-50, (2008)
- [Huang03] Huang, H.; Barker, W. C.; Chen, Y.; Wu, C. H.: iProClass: an integrated database of protein family, function and structure information.
Nucleic Acids Res., 31, 390-392, (2003)

- [Hubbard97] Hubbard, T. J. P.; Murzin, A. G.; Brenner, S. E.; Chothia, C.: SCOP: a Structural Classification of Proteins database.
Nucleic Acids Res., 25, 236-239, (1997)
- [Hunter09] Hunter, S.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R. D.; Gough, J.; Haft, D.; Hulo, N.; Kahn, D.; Kelly, E.; Laugraud, A.; Letunic, I.; Lonsdale, D.; Lopez, R.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; Mistry, J.; Mitchell, A.; Mulder, N.; Natale, D.; Orengo, C.; Quinn, A. F.; Selengut, J. D.; Sigrist, C. J.; Thimma, M.; Thomas, P. D.; Valentin, F.; Wilson, D.; Wu, C.H.; Yeats, C.: InterPro: the integrative protein signature database.
Nucleic Acids Res., 37, D211-D215, (2009)
- [Hutchinson93] Hutchinson, E. G., Thornton, J. M.: The Greek key motif: extraction, classification and analysis.
Prot. Eng., 6, 233-245, (1993)
- [Huysmans91] Huysmans, M.; Richelle, J.; Wodak, S. J.: SESAM: A Relational Database for Structure and Sequence of Macromolecules.
Proteins, 11, 59-76, (1991)
- [Hühne07] Hühne, R.; Koch, F.-T.; Sühnel, J.: A comparative view at comprehensive information resources on three-dimensional structures of biological macro-molecules.
Briefings in functional genomics and proteomics, (2007)
- [Ilari08] Ilari, A.; Savino, C.: Protein Structure Determination by X-Ray Crystallography.
Methods Mol. Biol., 452, 63-87, (2008)
- [Islam89] Islam, S. A.; Sternberg, M. J. E.: A relational database of protein structures designed for flexible enquiries about conformation.
Protein Eng., 2, 431-442, (1989)
- [Jaya07] Jaya, M. I.; Zainol, Z.; M.; Ahamed Hassain Malim, N. H.: iProt - A Data Warehouse for Protein Database.
Proceedings of the International Conference on Electronical Engineering and Informatics, 1040-1043, (2007)
- [Jupp88] Jupp, R. A.; Richards, A. D.; Kay, J.; Dunn, B. M.; Wyckoff, J. B.; Samloff, I. M.; Yamamoto, K.: Identification of the aspartic proteinases from human erythrocyte membranes and gastric mucosa (slow-moving proteinase) as catalytically equivalent to cathepsin E.
Biochem. J., 254, 895-898, (1988)
- [Kabsch76] Kabsch, W.: A solution for the best rotation to relate two sets of vectors.
Acta Cryst., A32, 922-923, (1976)
- [Kabsch83] Kabsch, W.; Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.
Biopolymers, 22, 2577-2637, (1983)
- [Kageyama98] Kageyama, T.: Molecular cloning, expression and characterization of an *Ascaris* inhibitor for pepsin and cathepsin E.
Eur. J. Biochem., 253, 804-809, (1998)
- [Karp72] Miller, R. E.; Thatcher, J. W.: Reducibility among combinatorial problems.
Complexity of Computer Computations: Proc. of a Symp. on the Complexity of Computer Computations, (The IBM Research Symposia Series, New York, NY: Plenum Press, 85-103) 1972

- [Kasuya99] Kasuya, A.; Thornton, J. M.: Three-dimensional Structure Analysis of PROSITE patterns.
J. Mol. Biol., 286, 1673-1691, (1999)
- [Keet03] Keet, M.: Biological Data and Conceptual Modelling Methods.
Journal of Conceptual Modeling, 29, (2003)
- [Keilova72] Keilova, H.; Tomasek, V.: EFFECT OF PEPSIN INHIBITOR FROM *ASCARIS LUMBRICOIDES* ON CATHEPSIN D AND E.
Biochim. Biophys. Acta, 284, 461-464, (1972)
- [Kelner94] Kelner, G. S.; Kennedy, J.; Bacon, K. B.; Kleyensteuber, S.; Largaespada, D. A.; Jenkins, N. A.; Copeland, N. G.; Bazan, J. F.; Moore, K. W.; Schall, T. J.; Zlotnik, A.: Lymphotactin: a novel cytokine which represents a new class of chemokine.
Science, 266, 1395-1399, (1994)
- [Kendrew58] Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckhoff, H.; Phillips, D. C.: A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.
Nature, 181, 662-6, (1958)
- [Kendrew60] Kendrew, J. C.; Dickerson, R. E.; Strandberg, B. E.; Hart, R. G.; Davies, D. R.; Phillips, D. C.; Shore, V. C.: Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution.
Nature, 185, 422-7, (1960)
- [Khafizov14] Khafizov, K.; Madrid-Aliste, C.; Almo, S. C.; Fiser, A.: Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative.
Proc. Natl. Acad. Sci., 111(10), 3733-8, (2014)
- [Kinjo17] Kinjo, A. R.; Bekker, G. J.; Suzuki, H.; Tsuchiya, Y.; Kawabata, T.; Ikegawa, Y.; Nakamura, H.: Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures.
Nucleic Acids Res., 45(D1), D282-D288, (2017)
- [Kiuchi07] Kiuchi, Y.; Ozaki, T.; Ohkawa, T.: Partial Geometric Hashing for Retrieving Similar Interaction Protein Using Profile.
Proceedings of the Fourth International Conference on Information Technology, ITNG07, (2007)
- [Knox07] Knox, D. P.: Proteinase inhibitors and helminth parasite infection.
Parasite Immunol., 29, 57-71, (2007)
- [Koch96] Koch, I.; Lengauer, T.; Wanke, E.: An Algorithm for Finding Maximal Common Subtopologies in a Set of Protein Structures.
J. Comp. Biol., 3, 289-306, (1996)
- [Kohlbacher00] Kohlbacher, O.; Lenhof, H.-P.: BALL - rapid software prototyping in computational molecular biology.
Bioinformatics, 16, 815-824, (2000)
- [Kolbeck06] Kolbeck, B.; May, P.; Schmidt-Goenner, T.; Steinke, T.; Knapp, E.-W.: Connectivity independent protein-structure alignment: a hierarchical approach.
BMC Bioinformatics, 7, 510, (2006)

- [Kologlu01] Kologlu, E. S.; Mccaslin, D. R.; Kitabwalla, M.; Pauza, C. D.; Markley, J. L.; Volkman, B. F.: Monomeric Solution Structure of the Prototypical 'C' Chemokine Lymphotactin. *Biochemistry*, 40, 12486-12496, (2001)
- [Köhler04] Köhler, J.: Integration of life sciences databases. *Biosilico*, 2, 61-69, (2004)
- [Krissinel04] Krissinel, E.; Henrick, K.: Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst., D60*, 2256-2268, (2004)
- [Krushna11] Krushna, N. S. A.; Shiny, C.; Manokaran, G.; Elango, S.; Narayanan, R. B.: Immune responses to recombinant *Brugia malayi* pepsin inhibitor homolog (Bm-33) in patients with human lymphotactic filariasis. *Parasitol. Res.*, 108, 407-415, (201)
- [Kufareva17] Kufareva, I.: Chemokines and their receptors: insights from molecular modeling and crystallography. *Curr. Opin. Pharmacol.*, 30, 27-37, (2016)
- [Labesse97] Labesse, G.; Colloc'h, N.; Pothier, J.; Mornon, J.-P.: P-SEA: a new efficient assignment of secondary structure from $C\alpha$ trace of proteins. *Comput. Appl. Biosci.*, 13, 291-295, (1997)
- [Lagerstrøm-Lang52] Lagerstroem-Lang, K. U.: Proteins and Enzymes. In: Lane Medical Lectures. Bd. 6, S. 1-115. Stanford University Publications, University Series, Medical Sciences, Stanford University Press, (1952)
- [Landsman09] Landsman, D.; Gentleman, R.; Kelso, J.; Ouellette, B. F. F.: DATABASE: A new forum for biological databases and curation. *Database*, 2009, bap002, (2009)
- [Lawrence01] Lawrence, R.: Biocomputing: impact of the genomic revolution. *DDT*, 6, 403-405, (2001)
- [Lawson08] Lawson, C. L.; Dutta, S.; Westbrook, J. K.; Henrick, K.; Berman, H. M.: Representation of viruses in the remediated PDB archive. *Acta Cryst., D64*, 874-882, (2008)
- [Leid87a] Leid, R. W.; Suquet, C. M.; Tanigoshi, L.: Parasite defense mechanisms for evasion of host attack: A review.. *Vet. Parasitol.*, 25, 147-162, (1987)
- [Leid87b] Leid, R. W.; Grant, R. F.; Suquet, C. M.: Inhibition of equine neutrophil chemotaxis and chemokinesis by a *Taenia taeniaeformis* proteinase inhibitor, taeniaestatin. *Parasite Immunol.*, 9, 195-204, (1987)
- [Lesk81] Lesk, A. M.; Rose, G. D.: Folding units in globular proteins. *Proc. Natl. Acad. Sci.*, 78, 4304-4308, (1981)
- [Leszczynski86] Leszczynski, J. F.; Rose, G. D.: Loops in Globular Proteins: A Novel Category of Secondary Structure. *Sci*, 234, 849-855, (1986)

- [Li00] Li, M.; Phylip, L. H.; Lees, W. E.; Winther, J. R.; Dunn, B. M.; Wlodawer, A.; Kay, J.; Gustchina, A.: The aspartic proteinase from *Saccharomyces cerevisiae* folds its own inhibitor into a helix.
Nat. Struct. Biol., 7, 113-117, (2000)
- [Lindqvist97] Lindqvist, Y.; Schneider, G.: Circular permutations of natural protein sequences: structural evidence.
Curr. Opin. Struct. Biol., 7, 422-427, (1997)
- [Luger97] Luger, K.; Mader, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J.: Crystal structure of the nucleosome core particle at 2.8Å resolution.
Nature, 389, 251-260, (1997)
- [Lynch08] Lynch, C.: How do your data grow?.
Nature (special issue: Big data), 455, 28-29, (2008)
- [Madej95] Madej, T.; Gibrat, J.-F.; Bryant, S. H.: Threading a Database of Protein Cores.
Proteins, 23, 356-369, (1995)
- [Maheshwari05] Maheshwari, A.; Lu, W.; Guida, W. C.; Christensen, R. D.; Calhoun, D. A.: IL-8/CXC Ligand 8 Survives Neonatal Gastric Digestion as a Result of Intrinsic Aspartyl Proteinase Resistance.
Pediatr. Res., 57, 438-444, (2005)
- [Majumdar05] Majumdar, I.; Krishna, S. S.; Grishin, N. V.: PALSSE: A program to delineate linear secondary structural elements from protein structures.
BMC Bioinformatics, 6, 202-225, (2005)
- [Mangalam02] Mangalam, H.: The Bio* toolkits - a brief overview.
Brief. Bioinform., 3, 296-302, (2002)
- [Martin05] Martin, J.; Letellier, G.; Marin, A.; Taly, J.-F.; de Breverén, A. G.; Gibrat, J.-F.: Protein secondary structure assignment revisited: a detailed analysis of different assignment methods.
BMC Struct. Biol., 5, 17-33, (2005)
- [Martzen90] Martzen, M. R.; McMullen, B. A.; Smith, N. E.; Fujikawa, K.; Peanasky, R. J.: Primary structure of the Major Pepsin Inhibitor from the Intestinal Parasitic Nematode *Ascaris suum*.
Biochemistry, 29, 7366-7372, (1990)
- [Mitchell89] Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P.: Use of Techniques Derived from Graph Theory to Compare Secondary Structure Motifs in Proteins.
J. Mol. Biol., 212, 151-166, (1989)
- [Mizuguchi95] Mizuguchi, K.; Go, N.: Comparison of spatial arrangements of secondary structural elements in proteins.
Prot. Eng., 8, 353-362, (1995)
- [Moose07] Moose, R. E.; Clemente, J. C.; Jackson, L. R.; Ngo, M.; Wooten, K.; Chang, R.; Bennett, A.; Chakraborty, S.; Yowell, C. A.; Dame, J. B.; Agbandje-McKenna, M.; Dunn, B. M.: Analysis of binding interactions of pepsin inhibitor-3 to mammalian and malarial aspartic proteases.
Biochemistry, 46(49):11198-205, (2007)
- [Morffew86] Morffew, A. J.; Todd, S. J. P.: The use of prolog as a protein query language.
Computers and Chemistry, 10, 9-14, (1986)

- [Murphy00] Murphy, P. M.; Baggiolini, M.; Charo, I. F.; Hebert, C. A.; Horuk, R.; Matsushima, K.; Miller, L. H.; Oppenheim, J. J.; Power, C. A.: International Union of Pharmacology. XXII. Nomenclature for Chemokine Receptors. *Pharmacol. Rev.*, 52, 145-176, (2000)
- [Murzin95] Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chotia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247, 536-540, (1995)
- [Murzin96] Murzin, A. G.: Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.*, 6, 386-394, (1996)
- [Nagano99] Nagano, N.; Ota, M.; Nishikawa, K.: Strong hydrophobic nature of cysteine residues in proteins. *FEBS Lett.*, 458, 69-71, (1999)
- [Nair09] Nair, R.; Liu, J.; Soong, T.-T.; Acton, T. B.; Everett, J. K.; Kouranov, A.; Fiser, A.; Godzik, A.; Jaroszewski, L.; Orengo, C.; Montelione, G. T.; Rost, B.: Structural genomics is the largest contributor of novel structural leverage. *J. Struct. Funct. Genomics*, 10, 181-191, (2009)
- [Ng00] Ng, K. K. S.; Petersen, J. F. W.; Cherney, M. M.; Garen, C.; Zalatoris, J. J.; Rao-Naik, C.; Dunn, B. M.; Martzen, M. R.; Peanasky, R. J.; James, M. N. G.: Structural basis for the inhibition of porcine pepsin by *Ascaris* pepsin inhibitor-3. *Nat. Struct. Biol.*, 7, 653-657, (2000)
- [Nietlispach04] Nietlispach, D.; Mott, H. R.; Stott, K. M.; Nielsen, P. R.; Thiru, A.; Laue, E. D.: Structure Determination of Protein Complexes by NMR. *Methods Mol. Biol.*, 278, 255-288, (2004)
- [Niiler01] Niiler, E.: Supercomputer for proteomics. *Nat. Biotechnol.*, 19, 193, (2001)
- [Niskanen03] Niskanen, S.; Østergård, P. R. J.: Cliquer User's Guide, Version 1.0. Communications Laboratory, Helsinki University of Technology, Espoo, Finland, Tech. Rep. T48, (2003)
- [Orengo97] Orengo, C. A.; Mitchie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M.: CATH - a hierarchic classification of domain structures. *Structure*, 5, 1093-1108, (1997)
- [Orengo99] Orengo, C. A.; Pearl, F. M. G.; Bray, J. E.; Todd, A. E.; Martin, A. C.; Lo Conte, L.; Thornton, J. M.: The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, 27, 275-279, (1999)
- [Ouzounis03] Ouzounis, C. A.; Coulson, R. M. R.; Enright, A. J.; Kunin, V.; Pereira-Leal, J. B.: Classification schemes for protein structure and function. *Nat. Rev. Genet.*, 4, 508-519, (2003)
- [Painter04] Painter, J.; Merritt, E. A.: *mmLib* Python toolkit for manipulating annotated structural models of biological macromolecules. *J. Appl. Cryst.*, 37, 174-178, (2004)

- [Peitsch95] Peitsch, M. C.; Stampf, D. R.; Wells, T. N. C.; Sussmann, J. L.: The Swiss-3DImage collection and PDB-Browser on the World-Wide Web.
TIBS, 20, 82-84, (1995)
- [Pennec94] Pennec, X.; Ayache, N.: An $O(n^2)$ Algorithm for 3D Substructure Matching of Proteins. Shape and Pattern Matching in Computational Biology, Plenum Publisher, (1994)
- [Pennec98] Pennec, X.; Ayache, N.: A geometric algorithm to find small but highly similar 3D substructures in proteins.
Bioinformatics, 14, 516-522, (1998)
- [Perutz60] Perutz, M. F.; Rossmann, M. G.; Cullis, A. F.; Muirhead, H.; Will, G.; North, A. C.: Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis.
Nature, 1960, 185, 416-22, (1960)
- [Pisabarro06] Pisabarro, M. T.; Leung, B.; Kwong, M. Corpuz, R.; Frantz, G. D.; Chiang, N.; Vandlen, R.; Diehl, L. J.; Skelton, N.; Kim, H. S.; Eaton, D.; Schmidt, K. N.: Cutting Edge: Novel Human Dendritic Cell- and Monocyte-Attracting Chemokine-Like Protein Identified by Fold Recognition Methods.
J. Immunol., 176, 2069-73, (2006)
- [Pitt01] Pitt, W. R.; Williams, M. A.; Steven, M.; Sweeney, B.; Bleasby, A. J.; Moss, D. S.: The Bioinformatics Template Library - generic components for biocomputing.
Bioinformatics, 17, 729-737, (2001)
- [Pocock00] Pocock, M.; Down, T.; Hubbard, T.: BioJava: open source components for bioinformatics.
ACM SIGBIO Newl., 20, 10-12, (2000)
- [Polat16] Polat, A.; Tunc, T.; Erdem, G.; Yerebasmaz, N.; Tas, A.; Beken, S.; Basbozkurt, G.; Saldır, M.; Zenciroglu, A.; Yaman, H.: Interleukin-8 and Its Receptors in Human Milk from Mothers of Full-Term and Premature Infants.
Breastfeeding Medicine, 11, 247-251, (2016)
- [Proudfoot04] Proudfoot, L.: Parasitic helminths tip the balance: potential anti-inflammatory therapies.
Immunology, 113, 438-440, (2004)
- [Przytycka99] Przytycka, T.; Aurora, R.; Rose, G. D.: A protein taxonomy based on secondary structure.
Nat. Struct. Biol., 6, 672-682, (1999)
- [Quinn15] Quinn, G. B.; Bi, C.; Christie, C. H.; Pang, K.; Prlic, A.; Nakane, T.; Zardecki, C.; Voigt, M.; Berman, H. M.; Bourne, P. E.; Rose, P. W.: RCSB PDB Mobile: iOS and Android mobile apps to provide data access and visualization to the RCSB Protein Data Bank.
Bioinformatics, 31(1), 126-7, (2015)
- [Rajaratnam99] Rajaratnam, K.; Sykes, B. D.; Dewald, B.; Baggiolini, M.; Clark-Lewis, I.: Disulfide Bridges in Interleukin-8 Probed Using Non-Natural Disulfide Analogues: Dissociation of Roles in Structure from Function.
Biochemistry, 38, 7653-7658, (1999)

- [Rajarathnam01] Rajarathnam, K.; Li, Y.; Rohrer, T.; Gentz, R.: Solution Structure and Dynamics of Myeloid Progenitor Inhibitory Factor-1 (MPIF-1), A Novel Monomeric CC Chemokine. *The Journal of Biological Chemistry*, 276, No. 7, 4909-4916, (2001)
- [Ram04] Ram, S.; Wei, W.: Modeling the Semantics of 3D Protein Structures. *The 23rd International Conference on Conceptual Modeling, (ER), Lecture Notes in Computer Science* 3288, 696-708, (2004)
- [Raman11] Raman, D.; Sobolik-Delmaire, T., Richmond, A.: Chemokines as targets for therapy. *Exp. Cell Res.*, 317(5), 602-12, (2011)
- [Ribeiro85] Ribeiro, J. M.; Makoul, G. T.; Levine, J.; Robinson, D. R.; Spielma, A.: Antihemostatic, antiinflammatory, and immunosuppressive properties of the saliva of a tick, *Ixodes dammini*. *J. Exp. Med.*, 161, 332-344, (1985)
- [Rigden16] Rigden, D. J.; Fernandez-Suarez, X. M.; Galperin, M. Y.: The 2016 Database Issue of Nucleic Acids Research and an updated Molecular Biology Database Collection. *Nucleic Acids Res.*, 44, D1-6, (2016)
- [Rigden18] Rigden, D. J.; Fernandez-Suarez, X. M.: The 2018 Nucleic Acids Research database issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.*, 44, D1-6, (2016)
- [Rose11] Rose, P. W.; Beran, B.; Bi, C.; Bluhm, W. F., Dimitropoulos, D.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D.; Young, J.; Yukich, B.; Berman, H. M.; Bourne P. E.: The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, 39 (Database issue), D392-401, (2011)
- [Rother04] Rother, K.; Müller, H.; Trissl, S.; Koch, I.; Steinke, T.; Preissner, R.; Frömmel, C.; Leser, U.: COLUMBA: Multidimensional Data Integration of Protein Annotations. *Lecture Notes in Computer Science*, 2994, 156-171, (2004)
- [Sali98] Sali, A.: 100,000 protein structures for the biologist. *Nat. Struct. Biol.*, 5, 1029-1032, (1998)
- [Samloff87] Samloff, I. M.; Taggart, R. T.; Shiraishi, T.; Branch, T.; Reid, W. A.; Heath, R.; Lewis, R. W.; Valler, M. J.; Kay, J.: Slow moving proteinase. Isolation, characterization, and immunohistochemical localization in gastric mucosa. *Gastroenterology*, 93, 77-84, (1987)
- [Sander91] Sander, C.; Schneider, R.: Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment. *Proteins*, 9, 56-68, (1991)
- [Scheeff03] Scheeff, E. D.; Fink, J. L.: Fundamentals of protein structures. *Methods Biochem. Anal.*, 44, 15-39, (2004)
- [Schneider96] Schneider, R.; Sander, C.: The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, 24, 201-205, (1996)
- [Schneider97] Schneider, R.; de Daruvar, A.; Sander, C.: The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, 25, 226-230, (1997)

- [Semertzidis93] Semertzidis, M. T.; Deplanque, S.; Mornon, J.-P.: OMEGA: a three-dimensional databank for protein structures (a complement to PDB).
Comput. Appl. Biosci., 9, 667-670, (1993)
- [Shabo06] Shabo, A.; Rabinovici-Cohen, S.; Vortman, P.: Revolutionary impact of XML on biomedical information interoperability.
IBM Systems Journal, 45, 361-372, (2006)
- [Shah05] Shah, S. P.; Huang, Y.; Xu, T.; Yuen, M. M. S.; Ling, J.; Ouellette, B. F. F.: Atlas - a data warehouse for integrative bioinformatics.
BMC Bioinformatics, 6, 34-49, (2005)
- [Shindyalov94] Shindyalov, I. N.; Chang, W.; Pu, C.; Bourne, P. E.: Macromolecular query language (MMQL): prototype data model and implementation.
Protein Eng., 7, 1311-1322, (1994)
- [Shindyalov95] Shindyalov, I. N.; Bourne, P. E.: WPDB - PC Windows based interrogation of macromolecular structure.
J. Appl. Cryst., 28, 847-852, (1995)
- [Shindyalov97] Shindyalov, I. N.; Bourne, P. E.: Protein data representation and query using optimized data decomposition.
Comput. Appl. Biosci., 13, 487-496, (1997)
- [Sillitoe15a] Sillitoe, I.; Lewis, T. E.; Cuff, A.; Das, S.; Ashford, P.; Dawson, N. L.; Furnham, N.; Laskowski, R. A.; Lee, D.; Lees, J. G.; Lehtinen, S.; Studer, R. A.; Thornton, J.; Orengo, C. A.: CATH: comprehensive structural and functional annotations for genome sequences.
Nucleic Acids Res., 43 (Database issue), D376-81, (2015)
- [Sillitoe15b] Sillitoe, I.; Dawson, N.; Thornton, J.; Orengo, C.: The history of the CATH structural classification of protein domains.
Biochimie, 119, 209-17, (2015)
- [Skolnick00] Skolnick, J.; Fetrow, J. S.; Kolinski, A.: Structural genomics and its importance for gene function analysis.
Nat. Biotechnol., 18, 283-287, (2000)
- [Stajich02] Stajich, J. E.; Block, D.; Boulez, K.; Brenner, S. E.; Chervitz, S. A.; Dagdigian, C.; Fuellen, G.; Gilbert, J. G. R.; Korf, Ian, Lapp, H.; Lehtinen, S.; Matsalla, C.; Mungall, C. J.; Osborne, B. I.; Pocock, M. R.; Schattner, P.; Senger, M.; Stein, L. D.; Stupka, E.; Wilkinson, M. D.; Birney, E.: The Bioperl Toolkit: Perl Modules for the Life Sciences.
Genome Res., 12, 1611-1618, (2002)
- [Stampf95] Stampf, D. R.; Felder, C. E.; Sussmann, J. L.: PDBBrowse - a graphics interface to the Brookhaven Protein Data Bank.
Nature, 374, 572-574, (1995)
- [Standley08] Standley, D. M.; Kinjo, A. R.; Kinoshita, K.; Nakamura, H.: Protein structure databases with new web services for structural biology and biomedical research.
Brief. Bioinform., 9, 276-285, (2008)
- [Stevens01] Stevens, R.; Goble, C.; Baker, P.; Brass, A.: A classification of tasks in bioinformatics.
Bioinformatics, 17, 180-188, (2001)

- [Stewart93] Stewart, G. W.: On the Early History of the Singular Value Decomposition.
SIAM Review, 35, 551-566, (1993)
- [Sunyaev05] Sunyaev, S.: Biocomputing enters its adolescence.
Genome Biol., 6, 325, (2005)
- [Sussmann98] Sussmann, J. L.; Jiang, D. L. J.; Manning, N. O.; Prilusky, J.; Ritter, O.; Abola, E. E.:
Protein Data Bank (PDB): Database of Three-dimensional Structural Information of Biological
Macromolecules.
Acta Cryst., D54, 1078-1084, (1998)
- [Sussmann99] Sussmann, J. L.; Abola, E. E.; Li, D.; Jiang, J.; Manning, N. O.; Prilusky, J.: The
protein data bank.
Genetica, 106, 149-158, (1999)
- [Szecsi92] Szecsi, P. B.: The aspartic proteases.
Scand. J. Clin. Lab. In vest. Suppl., 210, 5-22, (1992)
- [Thangudu08] Thangudu, R. R.; Manoharan, M.; Srinivasan, N; Cadet, F.; Sowdhamini, R.; Off-
mann, B.: Analysis on conservation of disulphide bonds and their structural features in homolo-
gous protein domain families.
BMC Struct. Biol., 8, 55-76, (2008)
- [Thornton81] Thornton, J. M.: Disulphide Bridges in Globular Proteins.
J. Mol. Biol., 151, 261-287, (1981)
- [Todd84] Todd, S.; Morffew, A.; Burrige, J.: Application of relational database and graphics to the
molecular sciences.
british computer society third british national conference on databases, 1-14, (1984)
- [Todd05] Todd, A. E.; Marsden, R. L.; Thornton, J. M.; Orengo, C. A.: Progress of Structural Ge-
nomics Initiatives: An Analysis of Solved Target Structures.
J. Mol. Biol., 348, 1235-1260, (2005)
- [Tramontano05] Tramontano A.: The ten most wanted solutions in protein bioinformatics.
Chapman & Hall / CRC Mathematical Biology and Medicine Series, (2005)
- [Trissl05] Trissl, S.; Rother, K.; Müller, H.; Steinke, T.; Koch, I.; Preissner, R.; Frömmel, C.; Leser,
U.: Columba: an integrated database of proteins, structures, and annotations.
BMC Bioinformatics, 6, 81-91, (2005)
- [Valler85] Valler, M. J.; Kay, J.; Aoyagi, T.; Dunn, B. M.: The interaction of aspartic proteinases
with naturally-occurring inhibitors from actinomycetes and *Ascaris lumbricoides*.
J. Enzyme Inhib., 1, 77-82, (1985)
- [von Adrian03] von Adrian, U. H.: Introduction: chemokines - regulation of immune cell trafficking
and lymphoid organ architecture.
Seminars in immunology, 15, 239-241, (2003)
- [Wadman99] Wadman, M.: Biologists make plea to NIH to invest in supercomputing centre.
Nature, 398, 93-94, (1999)
- [Wallace97] Wallace, A. C.; Borkakoti, N.; Thornton, J. M.: TESS: A geometric hashing algorithm
for deriving 3D coordinate templates for searching structural databases. Application to enzyme
active sites.
Protein Sci., 6, 2308-2323, (1997)

- [Wang05] Wang, X.; Gorlitsky, R.; Almeida, J. S.: From XML to RDF: how semantic web technologies will change the design of 'omic' standards.
Nat. Biotechnol., 23, 1099-1103, (2005)
- [Warr08] Warr, W. A.: Future structural genomics initiatives: an interview with Helen Berman, director of the Protein Data Bank.
J. Comput. Aided Mol. Des., 22, 707-710, (2008)
- [Weh02] Weh, M.: Anpassung von BLAST für Genom-Datenbanken.
diploma thesis, Humboldt Universität zu Berlin, chair databases and information systems, (2002)
- [Weissig98] Weissig, H.; Shindyalov, I. N.; Bourne, P. E.: Macromolecular Structure Databases: Past Progress and Future Challenges.
Acta Cryst., D54, 1085-1094, (1998)
- [Weskamp04] Weskamp, N.; Kuhn, D.; Hüllermeier, E.; Klebe, G.: Efficient similarity search in protein structure database by k-clique hashing.
Bioinformatics, 20, 1522-1526, (2004)
- [Westbrook02] Westbrook, J.; Feng, Z.; Jain, S.; Bhat, T. N.; Thanki, N.; Ravichandran, V.; Gilliland, G. L.; Bluhm, W. F.; Weissig, H.; Greer, D. S.; Bourne P. E.; Berman, H. M.: The Protein Data Bank: unifying the archive.
Nucleic Acids Res., 30, 245-248, (2002)
- [Westbrook05] Westbrook, J. D.; Ito, N.; Nakamura, H.; Henrick, K.; Berman, H. M.: PDBML: the representation of archival macromolecular structure data in XML.
Bioinformatics, 21, 988-992, (2005)
- [Wikel86] Wikel, S. K.; Whelen, A. C.: Ixodid-host immune interaction. Identification and characterization of relevant antigens and tick-induced host immunosuppression.
Vet. Parasitol., 20, 149-174, (1986)
- [Willenbücher93] Willenbücher, J.; Höfle, W.; Lucius, R.: The filarial antigens Av33/Ov33-3 show striking similarities to the major pepsin inhibitor from *Ascaris suum*.
Mol. Biochem. Parasitol., 57, 349-352, (1993)
- [Willett95] Willett, P.: Searching for Pharmacophoric Patterns in Databases of Three-dimensional Chemical Structures.
J. Mol. Recognit., 8, 290-303, (1995)
- [Touw15] Touw, W. G.; Baakman, C.; Black, J.; te Beek, T. A. H.; Krieger, E.; Joosten, R. P.; Vriend, G.: A series of PDB-related databanks for everyday needs.
NAR, 43, database issue, D364-D8, (2015)
- [Wlodawer08] Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M.: Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures.
FEBS J., 275, 1-21, (2008)
- [Wolfson97] Wolfson, H. J.; Rigoutsos, I.: Geometric Hashing: An Overview.
IEEE Comput. Sci. Eng., 10-21, (1997)
- [Wu01] Wu, C. H.; Xiao, C.; Hou, Z.; Huang, H.; Barker, W. C.: iProClass: an integrated, comprehensive, and annotated protein classification database.
Nucleic Acids Res., 29, 52-54, (2001)

- [Wüthrich90] Wüthrich, K.: Protein Structure Determination in Solution by NMR Spectroscopy.
J. Biol. Chem., 22059-22062, (1990)
- [Yuan05] Yuan, X.; Bystroff C.: Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins.
Bioinformatics, 21, 1010-1019, (2005)
- [Zalatoris99] Zalatoris, J. J.: Development and partial structural characterization of a recombinant inhibitor of pepsin from *Ascaris*.
Ph. D. Thesis, University of Florida, (1999)
- [Zlotnik00] Zlotnik, A.; Yoshie, O.: Chemokines: a new classification system and their role in immunity.
Immunity, 12, 121-127, (2000)

The Intrinsic Pepsin Resistance of Interleukin-8 Can be Explained from a Combined Bioinformatical and Experimental Approach

Gerd Anders, Ulrich Hassiepen, Stephan Theisgen, Stephan Heymann, Lionel Muller,
Tania Panigada, Daniel Huster, Sergey A. Samsonov

Abstract—Interleukin-8 (IL-8, CXCL8) is a neutrophil chemotactic factor belonging to the family of chemokines. IL-8 was shown to resist pepsin cleavage displaying its high resistance to this protease. However, the molecular mechanisms underlying this resistance are not fully understood. Using our in-house database containing the data on three-dimensional arrangements of secondary structure elements from the whole Protein Data Bank, we found a striking structural similarity between IL-8 and pepsin inhibitor-3. Such similarity could play a key role in understanding IL-8 resistance to the protease pepsin. To support this hypothesis, we applied pepsin assays confirming that intact IL-8 is not degraded by pepsin in comparison to IL-8 in a denaturated state. Applying ^1H - ^{15}N Heteronuclear Single Quantum Coherence NMR measurements, we determined the putative regions at IL-8 that are potentially responsible for interactions with the pepsin. The results obtained in this work contribute to the understanding of the resistance of IL-8 to pepsin proteolysis in terms of its structural properties.

Index Terms— Protein-protein interactions, computational biology, pepsin resistance, pepsin inhibitor, interleukin-8

With the signatures below we confirm, that our contribution of the aforementioned publication is part of the experimental approaches, but not part of any described bioinformatical approach:

Ulrich Hassiepen:

.....

Stephan Theisgen:

.....

Stephan Heymann:

Lionel Muller:

.....

Tania Panigada:

.....

Daniel Huster:

Sergey A. Samsonov:

The Intrinsic Pepsin Resistance of Interleukin-8 Can be Explained from a Combined Bioinformatical and Experimental Approach

Gerd Anders, Ulrich Hassiepen, Stephan Theisgen, Stephan Heymann, Lionel Muller, Tania Panigada, Daniel Huster, Sergey A. Samsonov

Abstract—Interleukin-8 (IL-8, CXCL8) is a neutrophil chemotactic factor belonging to the family of chemokines. IL-8 was shown to resist pepsin cleavage displaying its high resistance to this protease. However, the molecular mechanisms underlying this resistance are not fully understood. Using our in-house database containing the data on three-dimensional arrangements of secondary structure elements from the whole Protein Data Bank, we found a striking structural similarity between IL-8 and pepsin inhibitor-3. Such similarity could play a key role in understanding IL-8 resistance to the protease pepsin. To support this hypothesis, we applied pepsin assays confirming that intact IL-8 is not degraded by pepsin in comparison to IL-8 in a denaturated state. Applying ^1H - ^{15}N Heteronuclear Single Quantum Coherence NMR measurements, we determined the putative regions at IL-8 that are potentially responsible for interactions with the pepsin. The results obtained in this work contribute to the understanding of the resistance of IL-8 to pepsin proteolysis in terms of its structural properties.

Index Terms— Protein-protein interactions, computational biology, pepsin resistance, pepsin inhibitor, interleukin-8

With the signatures below we confirm, that our contribution of the aforementioned publication is part of the experimental approaches, but not part of any described bioinformatical approach:

Ulrich Hassiepen:

Stephan Theisgen:

.....

Stephan Heymann:

.....

Lionel Muller:

Tania Panigada:

Daniel Huster:

.....

Sergey A. Samsonov:

.....

Anhang A

Erklärung über die eigenständige Abfassung der Arbeit

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne unzulässige Hilfe oder Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Ich versichere, dass Dritte von mir weder unmittelbar noch mittelbar eine Vergütung oder geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen, und dass die vorgelegte Arbeit weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zweck einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt wurde. Alles aus anderen Quellen und von anderen Personen übernommene Material, das in der Arbeit verwendet wurde oder auf das direkt Bezug genommen wird, wurde als solches kenntlich gemacht. Insbesondere wurden alle Personen genannt, die direkt an der Entstehung der vorliegenden Arbeit beteiligt waren. Die aktuellen gesetzlichen Vorgaben in Bezug auf die Zulassung der klinischen Studien, die Bestimmungen des Tierschutzgesetzes, die Bestimmungen des Gentechnikgesetzes und die allgemeinen Datenschutzbestimmungen wurden eingehalten. Ich versichere, dass ich die Regelungen der Satzung der Universität Leipzig zur Sicherung guter wissenschaftlicher Praxis kenne und eingehalten habe.

München, 10.02.2018

.....
Ort, Datum

Gerd Anders

.....
Gerd Anders

Anhang B

Beruflicher Werdegang

Anhang C

Publikationen

Publikationen als Erstautor:

- **Anders, G.**; Hassiepen, U.; Theisgen, S.; Heymann, S.; Muller, L.; Panigada, T.; Huster, D.; Samsonov, S. A.
The Intrinsic Pepsin Resistance of Interleukin-8 Can be Explained from a Combined Bioinformatical and Experimental Approach
IEEE/ACM Trans. Comput. Biol. Bioinform., PP(99), 2016
- **Anders, G.**; Mackowiak, S. D.; Jens, M.; Maaskola, J.; Kuntzagk, A.; Rajewsky, N.; Landthaler, M.; Dieterich, C.
doRiNA: a database of RNA interactions in post-transcriptional regulation
NAR, 40(Database issue): D180-D186, 2012
- **Anders, G.**; Nicola, M.
Managing the Protein Data Bank with DB2 pureXML
IBM developerworks, 2011

Sonstige Publikationen:

- Samsonov, S. A.; Salwiczek, M.; **Anders, G.**; Koksche, B.; Pisabarro, M. T.
Fluorine in Protein Environments: A QM and MD Study
J. Phys. Chem. B., 113(51), 16400-8, 2009
- Samsonov, S. A.; Teyra, J.; **Anders, G.**; Pisabarro, M. T.
Analysis of the impact of solvent on contacts prediction in proteins
BMC Struct. Biol., 9:22, 2009
- Teyra, J.; Paszkowski-Rogacz, M.; **Anders, G.**; Pisabarro, M. T.
SCOWLP classification: Structural comparison and analysis of protein binding regions
BMC Bioinformatics, 9:9, 2008

Anhang D

Danksagung

An dieser Stelle möchte ich mich bei Familienangehörigen, KollegInnen, Freunden ganz herzlich bedanken, die mich auf vielfältige Weise unterstützt haben. Die folgende Liste erhebt keinen Anspruch auf Vollständigkeit (und ist alphabetisch sortiert):

- Anders†, G. (OMR Prof. Dr.)
- Anders, M.
- Dieterich, Ch. (Dr.)
- Freytag, J.-Ch. (Prof. Ph. D.)
- Hassiepen, U. (Dr.)
- Heymann, S. (Dr.)
- Huster, D. (Prof. Dr.)
- Loeser, H. (Dr.)
- Nicola, M. (Dr.)
- Pisabarro, M. T. (Dr.)
- Samsonov, S. A. (Dr.)
- Teyra, J. (Dr.)
- Tschira†, K. (Dr.)
- Zelazny, B. (Dr.)
- Zelazny, Ch. (Dr.)

Diese Dissertation ist nicht nur, aber insbesondere Prof. Dr. med. G. Anders gewidmet.