Biomolekulare Wechselwirkungen von Proteinen – Datenbankunterstützung für das Proteindocking

Abschlußbericht BIOWEPRO-Teilvorhaben München[1]

Hans-Peter Kriegel, Thomas Schmidt, Thomas Seidl

Institut für Informatik, Universität München
Oettingenstr. 67, D-80538 München, Germany
email: kriegel@dbs.informatik.uni-muenchen.de
http://www.dbs.informatik.uni-muenchen.de

Zusammenfassung

Die bislang bekannten Methoden zum 1:1-Docking eignen sich aus technischen und qualitativen Gründen nur wenig zum Einsatz für die 1:n-Dockingsuche. Insbesondere mangelt es an der Effizienz beim Einsatz für große Proteindatenbanken, wofür neue Methoden zur Unterstützung der Dockingsuche benötigt werden. Ziel des Vorhabens ist es, effiziente Verfahren zur Ähnlichkeitssuche in Proteindatenbanken zu entwickeln, die sich als Filterschritt für die 1:n-Dockingvorhersage eignen. Als wichtiger neuer Ansatz wurde das approximationsbasierte Ähnlichkeitsmaß für 3D-Oberflächensegmente entwickelt, das bereits erfolgreich auf Dockingsegmente in einer Proteindatenbank angewendet wurde. Zur effizienten Bearbeitung der zugehörigen Anfragen wurde die Ellipsoidanfrage als ein neuer Anfragetyp identifiziert, der eine breite und flexible Anwendbarkeit für die Ähnlichkeitssuche in Proteindatenbanken bietet. Unsere neuen Algorithmen unterstützen die effiziente Bearbeitung von Ellipsoidanfragen mit Hilfe mehrdimensionaler Indexstrukturen. Insbesondere für den Filterschritt beim 1:n-Proteindocking zeigen sich vielversprechende Ergebnisse. Weitere Anwendungen dieser Methoden erstrecken sich auf große Molekül-, Multimedia- und CAD-Datenbanken.

Dieser Abschlußbericht stellt den Sachbericht des Verwendungsnachweises im Sinne von Nr. 6.2 ANBest-P dar. Während Abschnitt 1 eine kurze Darstellung der Rahmenbedingungen gemäß Nr. 5b.2.1 BNBest enthält, wird in den Abschnitten 2 und 3 das erzielte Ergebnis nach Nr. 5b.2.2 BNBest eingehend dargestellt. Abschnitt 4 beinhaltet die Angaben über die verwendete Literatur und die Veröffentlichungen des Projektergebnisses gemäß Nr. 9a.3 BNBest.

Inhalt

---

1. Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung und Technologie unter dem Förderkennzeichen 01 IB 307 B0 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart
	Schlußbericht

3a. Titel des Berichts

Biomolekulare Wechselwirkungen von Proteinen – Datenbankunterstützung für das Proteindocking. Abschlußbericht BIOWEPRO-Teilvorhaben München

3b. Titel der Publikation

4a. Autoren des Berichts (Name, Vorname(n))	5. Abschlußdatum des Vorhabens
Kriegel, Hans-Peter; Schmidt, Thomas; Seidl, Thomas	Mai 1997
	6. Veröffentlichungsdatum: Dezember 1997
4b. Autoren der Publikation (Name, Vorname(n))	7. Form der Publikation
8. Durchführende Institution(en) (Name, Adresse) Lehr- und Forschungseinheit für Datenbanksysteme Institut für Informatik Universität München Oettingenstr. 67 D-80538 München	9. Ber.Nr. Durchführende Institution
	10. Förderkennzeichen: 01 IB 307 B0
	11a. Seitenzahl Bericht: 14
	11b. Seitenzahl Publikation
13. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) 53170 Bonn	12. Literaturangaben: 46
	14. Tabellen: 0
	15. Abbildungen: 2

16. Zusätzliche Angaben

17. Vorgelegt bei (Titel, Ort, Datum)

verschiedenen Fachtagungen (s. Literaturangaben)

18. Kurzfassung

Die bislang bekannten Methoden zum 1:1-Docking eignen sich aus technischen und qualitativen Gründen nur wenig zum Einsatz für die 1:n-Dockingsuche. Insbesondere mangelt es an der Effizienz beim Einsatz für große Proteindatenbanken, wofür neue Methoden zur Unterstützung der Dockingsuche benötigt werden. Ziel des Vorhabens ist es, effiziente Verfahren zur Ähnlichkeitssuche in Proteindatenbanken zu entwickeln, die sich als Filterschritt für die 1:n-Dockingvorhersage eignen. Als wichtiger neuer Ansatz wurde das approximationsbasierte Ähnlichkeitsmaß für 3D-Oberflächensegmente entwickelt, das bereits erfolgreich auf Dockingsegmente in einer Proteindatenbank angewendet wurde. Zur effizienten Bearbeitung der zugehörigen Anfragen wurde die Ellipsoidanfrage als ein neuer Anfragetyp identifiziert, der eine breite und flexible Anwendbarkeit für die Ähnlichkeitssuche in Proteindatenbanken bietet. Unsere neuen Algorithmen unterstützen die effiziente Bearbeitung von Ellipsoidanfragen mit Hilfe mehrdimensionaler Indexstrukturen. Insbesondere für den Filterschritt beim 1:n-Proteindocking zeigen sich vielversprechende Ergebnisse. Weitere Anwendungen dieser Methoden erstrecken sich auf große Molekül-, Multimedia- und CAD-Datenbanken.

19. Schlagwörter

Molekulare Bioinformatik; Vorauswahl zur 1:n-Dockingsuche; Ähnlichkeitssuche in 3D-Proteindatenbanken

20. Verlag	21. Preis

Document Control Sheet

1. ISBN or ISSN	2. Type of Report
	Final Report

3a. Report Title

Biomolecular Interactions of Protein – Database Support for the Protein Docking Problem. Final Report BIOWEPRO - Munich (in German)

3b. Title of Publication

4a. Author(s) of the Report (Family Name, First Name(s))	5. End of Project
Kriegel, Hans-Peter; Schmidt, Thomas; Seidl, Thomas	May 1997
	6. Publication Date: December 1997
4b. Author(s) of the Publication (Family Name, First Name(s))	7. Form of Publication
8. Performing Organization(s)	9. Originator's Report No.
Institute for Computer Science University of Munich Oettingenstr. 67 D-80538 Munich Germany	10. Reference No.: 01 IB 307 B0
	11a. No. of Pages Report: 14
	11b. No. of Pages Publication: 46
13. Sponsoring Agency (Name, Address)	12. No. of References: 46
Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) 53170 Bonn	14. No. of Tables: 0
	15. No. of Figures: 2

16. Supplementary Notes

17. Presented at (Title, Place, Date)

several conferences (see references)

18. Abstract

The presently available methods for the one-to-one docking problem are not well suited for the one-to-many docking problem for technical as well as quality reasons. In particular, there is a lack of efficiency when applying the methods to large protein databases. Therefore, new techniques to support the docking search are required. The project aims at the development of efficient methods for similarity search in protein database systems which are applicable as a filter step for the one-to-many docking prediction. As a new approach, we developed the approximation-based similarity measure for 3-D surface segments and successfully applied the model to docking segments in a protein database. For the efficient processing of similarity queries, we identified the ellipsoid query as a new query type which provides a wide and flexible applicability for similarity search in protein database systems. Our new algorithms support efficient processing of ellipsoid queries based on multidimensional index structures. In particular, applied to the filter step of one-to-many protein docking, promising results are obtained. Additional applications of the methods range from large molecular databases to multimedia and CAD database systems.

19. Keywords

Molecular Bioinformatics; Preselection for one-to-many docking search; Similarity search in 3-D protein database systems

20. Publisher	21. Price

1 Rahmenbedingungen

1.1 Aufgabenstellung

Das Verbundprojekt BIOWEPRO diente der Entwicklung von Algorithmen zur effizienten Beantwortung der Frage nach den möglichen Wechselwirkungen eines Proteins mit allen in einer Datenbank erfaßten Proteinen, deren räumliche Struktur bekannt ist. Dabei sollten insbesondere Lösungen zu Fragestellungen entwickelt werden wie etwa: Welche Proteine der DB kommen als Dockingkandidaten für das Anfrageprotein in Frage, an welchen Stellen findet die Wechselwirkung statt, und wie kann die Wechselwirkung im Hinblick auf ihre Stabilität bewertet werden?

Das Teilvorhaben der Arbeitsgruppe München beinhaltete die Entwicklung von neuen Datenbanktechniken zur Unterstützung der Dockingsuche auf großen Proteindatenbanken. Im einzelnen ergaben sich die folgenden Aufgaben:

(A) Auswahl von geeigneten Repräsentationen für Proteinoberflächen sowie Entwicklung von 3D-Formbeschreibungen.
(B) Entwicklung und Bereitstellung von Datenbank-Anfragemechanismen zum Auffinden von Dockingkandidaten und Dockingstellen.
(C) Entwicklung eines Nichtstandard-Datenbankmanagementsystems zur Identifikation biomolekularer Wechselwirkungen

Die Entwicklung einer geeigneten Repräsentationsform für die Oberflächen der Proteine war notwendig, da die Geometrie der Oberfläche für die Ausbildung von Wechselwirkungen eine entscheidende Rolle spielt. Die Repräsentation muß dabei für sehr große Datenbanken mit zehn- bis hunderttausenden von Molekülen geeignet sein, in denen eine effiziente Suche nach möglichen Dockingkandidaten durchgeführt werden soll. Basierend auf der Oberflächenrepräsentation sollten geeignete Anfragemechanismen entwickelt werden, durch die in einer groben Vorsuche die Kandidatenmenge möglichst weit eingeschränkt werden kann. Auf diese Weise müssen in nachgeschalteten, aufwendigeren Bearbeitungs- und Bewertungsschritten nur noch eine kleine Anzahl von Molekülen betrachtet werden.

1.2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Vorhaben (Entwicklung einer für Datenbanken geeigneten Repräsentationsform der Oberflächenbeschreibung) wurde im Rahmen des Verbundprojektes BIOWEPRO (Biomolekulare Wechselwirkungen von Proteinen) durchgeführt. Als Projektpartner arbeitete unsere Münchner Arbeitsgruppe (Prof. Dr. H.-P. Kriegel) mit der Gesellschaft für Biotechnologische Forschung (GBF) in Braunschweig (Prof. Dr. D. Schomburg, Projektkoordinator), der Technischen Fakultät der Universität Bielefeld (Prof. Dr. G. Sagerer) und dem Max-Planck-Institut für Biophysikalische Chemie in Göttingen (Dr. D. M. Soumpasis) zusammen.

Die Münchner Arbeitsgruppe konnte große Erfahrung in der Repräsentation, Speicherung und Suche von komplexen, geometrischen Objekten in das Projekt einbringen. Diese hatte sie durch langjähriges Arbeiten im Bereich Anfragemechanismen für Geo-Datenbanken, insbesondere über räumliche Indexstrukturen sowie Zerlegungs- und Approximationstechniken, erworben.

1.3 Planung und Ablauf des Vorhabens

Das Vorhaben wurde für eine zeitliche Dauer von drei Jahren für drei Bearbeiter (108 Bearbeitermonate, BM) geplant. Davon sollten 28 BM auf die Aufgabe A (Oberflächenrepräsenta-

tion), 36 BM auf die Aufgabe B (Datenbank-Anfragemechanismen) und 42 BM auf die Aufgabe C (Proteindatenbanksystem) entfallen. Da für das Projekt schließlich nur zwei Bearbeiter genehmigt wurden, konnten der ursprüngliche Arbeitsplan nicht in vollem Umfang erfüllt werden. So wurde die Entwicklung des Proteindatenbanksystems (Aufgabe C) zwar durchgeführt, doch war die geplante Integration verschiedener Module zu einem echten 1:n-Dockingsystem nicht mehr möglich. In Bezug auf die Oberflächenrepräsentation (Aufgabe A) und die Datenbank-Anfragemechanismen (Aufgabe B) konnten wir trotz der Einschränkung des Kostenrahmens sehr gute Forschungsergebnisse erzielen. Das Erreichen dieser Ziele wurde insbesondere durch die Verlängerung der Projektlaufzeit bis Ende Mai 1997 ermöglicht, die aufgrund eigener Haushaltsmittel kostenneutral bewerkstelligt werden konnte.

Die Bearbeitung des Projektes wurde in enger Zusammenarbeit mit den Projektpartnern durchgeführt. Neben dem fernmündlichen und email-gestützten Austausch wurden zahlreiche BIOWEPRO-Arbeitstreffen durchgeführt, die weiter unten einzeln aufgeführt werden.

1.4 Wissenschaftlicher und Technischer Ausgangspunkt

Auf dem Gebiet des Dockings gab es eine Reihe von Vorschlägen für das 1:1-Docking-Problem, d.h. für das Auffinden von möglichen Komplexkonformationen für zwei gegebene Proteine. Für das 1:n-Docking, also die Bestimmung von Dockingpartnern für ein gegebenes Protein aus einer großen Menge von Molekülen gab und gibt es bis heute keine vergleichbaren Ansätze in der Literatur.

Die Speicherung von Geo-Objekten in Datenbanksystemen, sowie die zugehörige Anfragebearbeitung ist ein Problem, das bereits gut untersucht war. Dies galt allerdings nur für den Fall von zweidimensionalen Objekten, z.B. Polygone, wie sie typischerweise bei geographischen Anwendungen auftreten. Die Anfragen auf diesen Objekten beschränkten sich auf lageabhängige Bedingungen, wie z.B. Schnitt mit anderen Objekten. Zur Speicherung von 3D-Objekten in Datenbanken, sowie zur Bearbeitung von formabhängigen Anfragen, wie sie beim Proteindocking auftreten, waren zu Projektbeginn weder eigene noch fremde Publikationen verfügbar.

1.5 Verwendete Konstruktionen und Verfahren

Für die Bearbeitung des Vorhabens wurde insbesondere die aus den Projektmitteln finanzierte Workstation HP-9000/735 verwendet. Als technische Basis für unser neues 3D-Proteindatenbanksystem kam das objektorientierte Datenbanksystem ObjectStore zum Einsatz. Neben diesen zugekauften Hilfsmitteln wurden insbesondere verschiedene mehrdimensionale Indextechniken verwendet, die vor und während der Projektlaufzeit unabhängig vom Projekt an unserer Lehr- und Forschungseinheit entwickelt worden sind. Hierbei sind konkret die Implementierungen des R*-Baumes [BKSS 90] und des X-Baumes [BKK 96] zu nennen.

1.6 Verwendete Fachliteratur

Die verwendete Fachliteratur ist zwei groben Bereichen zuzuordnen. Diese Gliederung liegt auch dem Literaturverzeichnis am Ende des Berichtes zugrunde. Zum einen wurden Publikationen ausgewertet, die bisherige Verfahren und Ergebnisse aus der Molekularen Bioinformatik darstellen. Darunter fallen Arbeiten zum Thema Proteine und Moleküloberflächen [Ber+ 77] [Con 83] [Con 86] [HO 94] [VBW 94] sowie bisherige Vorschläge zum 1:1-Dokkingproblem [AHPS 95] [BMH 92] [FNNW 93] [Kat+ 92] [SBK 92]. Der zweite Literaturbereich betrifft Datenbanktechniken wie effiziente Anfragetechniken in Geodatenbanksystemen

[BKSS 90] [BHKS 93] [BK 94] [BKK 96] [KSB 93] [OHMS 92] [Ore 86] sowie Ansätze zur geometrischen Ähnlichkeitssuche [AFS 93] [BS 96] [FRM 94] [HS 95] [Kor+ 96] [SKSH 89].

1.7 Zusammenarbeit mit anderen Stellen

Die Arbeiten im Projekt BIOWEPRO wurden in enger Kooperation mit den Projektpartnern in Braunschweig, Bielefeld und Göttingen durchgeführt. Auf den regelmäßig abgehaltenen Arbeitstreffen wurden jeweils die jüngsten Ergebnisse vorgestellt, neue Lösungsansätze diskutiert und das weitere Vorgehen koordiniert. Diese BIOWEPRO-Arbeitstreffen fanden sowohl als eigenständige Termine als auch in Verbindung mit Fachtagungen statt:

02.09.1993	BIOWEPRO-Arbeitstreffen in Braunschweig
10.12.1993	Arbeitstreffen mit Dr. Michael Meyer (Braunschweig) in München
19.01.1994	BIOWEPRO-Arbeitstreffen in München
10.03.1994	BIOWEPRO-Arbeitstreffen in Bielefeld
27.-28.06.1994	BIOWEPRO-Arbeitstreffen in Bielefeld
05.-07.09.1994	Tagungsbeitrag [Ald+ 94] und BIOWEPRO-Arbeitstreffen auf der 2. GI-Fachtagung *'Computeranwendungen in den Biowissenschaften'* in Jena.
11.11.1994	BIOWEPRO-Arbeitstreffen in Göttingen
22.-24.03.1995	Tagungsbeitrag [EKSX 95] auf der GI-Fachtagung *'Datenbanken in Büro, Technik und Wissenschaft'* (BTW'95) in Dresden.
12.05.1995	BIOWEPRO-Arbeitstreffen in Braunschweig
16.-19.07.1995	Tagungsbeitrag [SK 95a] und BIOWEPRO-Arbeitstreffen auf der *'International Conference on Intelligent Systems for Molecular Biology'* (ISMB-95) in Cambridge (England).
10.-14.07.1995	Vortrag auf dem *International GI-Seminar 'Molecular Bioinformatics'* in Dagstuhl, Saarland.
06.-09.08.1995	Vorträge [SK 95b] und [EKX 95a] auf dem *'International Symposium on Large Spatial Databases'* (SSD'95) in Portland, Maine.
20.-21.08.1995	Vortrag [EKX 95b] auf der *'International Conference on Knowledge Discovery and Data Mining'* (KDD'95) in Montreal.
22.09.1995	BIOWEPRO-Arbeitstreffen in München.
09.-11.10.1995	Tagungsbeitrag [KSS 95] und BIOWEPRO-Arbeitstreffen auf der GI-Fachtagung *'Bioinformatik – Informatik in den Biowissenschaften'* in Braunschweig.
07.-08.11.1995	Präsentation der Ergebnisse [Sch+ 95] und BIOWEPRO-Arbeitstreffen auf dem *BMBF-Statusseminar 'Molekulare Bioinformatik und Evolutionäre Algorithmen'* in Braunschweig.
26.-28.11.1995	Teilnahme am *'German-Israeli Symposium on Computer Science Aspects of Molecular Biology'*, Tel-Aviv University, Israel, mit Vortrag zum Thema *'Database Support for Protein Docking'*.
25.04.1996	BIOWEPRO-Arbeitstreffen in Bielefeld
20.09.1996	BIOWEPRO-Arbeitstreffen in Braunschweig
30.09.-02.10.1996	Tagungsbeitrag [EKW 96] und BIOWEPRO-Arbeitstreffen mit den Projektpartnern auf der GI-Fachtagung *'German Conference on Bioinformatics'* (GCB'96) in Leipzig

15.-18.07.1997 Tagungsbeitrag [KSS 97] auf dem *'International Symposium on Large Spatial Databases* (SSD'97) in Berlin.

25.-29.09.1997 Tagungsbeitrag [SK 97] auf der *'23rd International Conference on Very Large Data Bases'* (VLDB'97) in Athen, Griechenland.

2 Im Projekt erzieltes Ergebnis

In diesem Abschnitt beschreiben wir das in unserem Teilvorhaben 'Datenbankunterstützung für die Proteindockingsuche' erzielte Forschungsergebnis. Im einzelnen werden unser objektorientiertes Proteindatenbanksystem OS_PDB mit den darin enthaltenen Molekülоberflächen erläutert, das Basismodell für die mehrstufige Dockinganfragebearbeitung als Rahmen für die von uns entwickelten Filterschritte für die Dockingsuche, das approximationsbasierte Ähnlichkeitsmodell für 3D-Oberflächensegmente als ein konkreter Ansatz für einen geometrischen Filterschritt, sowie weiterführende Arbeiten, die auf dem im Projekt entwickelten allgemeinen und flexiblen Konzept der Ellipsoidanfrage beruhen.

2.1 Objektorientierte Proteindatenbank und Molekülоberflächen

Für das Docking benötigt man eine große Menge von Daten aus verschiedenen Quellen und in unterschiedlichen Formaten. Die Grundlage bilden Strukturdaten, wie sie in der Brookhaven Protein Data Bank (PDB) als Textfiles verfügbar sind. Da für die Ausbildung von Wechselwirkungen zwischen Proteinen jedoch die Komplementarität der Oberflächen eine wichtige Vorbedingung ist, ist die alleinige Darstellung der Proteinstruktur durch die Atomkoordinaten nicht ausreichend. Vielmehr ist eine Repräsentation der Molekülоberfläche notwendig, aus welcher Formbeschreibungen der Proteine, z.B. durch geometrische Kennzahlen, abgeleitet werden können. Darüber hinaus benötigt ein Dockingalgorithmus Zugriff auf physikochemische Parameter wie van-der-Waals-Radien oder Partialladungen.

Unser Proteindatenbanksystem OS_PDB [Som 94] basiert technisch auf dem kommerziellen objektorientierten Datenbanksystem ObjectStore [OHMS 92]. Sämtliche für das Docking notwendigen Proteindaten sind in OS_PDB integriert, und alle Vorverarbeitungsschritte, wie z.B. die Oberflächen- und Kennzahlberechnung stützen sich auf OS_PDB ab, d.h. sie erhalten ihre Eingabe aus der Datenbank und schreiben ihre Ausgabe wieder dorthin zurück.

Als Oberflächenmodell für Moleküle verwenden wir die Connolly-Oberfläche, die aus konvexen, sattelförmigen und konkaven Flächenstücken (Patches) besteht [Sch 94]. Zur Speicherung setzen wir die speziell für diesen Zweck entwickelte TriEdge-Datenstruktur ein [SK 95a] [SK 95b]. Dadurch wird einerseits eine besonders redundanzarme und damit speichereffiziente Darstellung erreicht, in der die hohe Regularität der Oberflächenstruktur ausgenutzt wird. Andererseits stehen eine Reihe von Operationen zur Verfügung, die ein Navigieren auf der Oberfläche ermöglichen.

Die analytische Repräsentation durch Patches bringt zudem den Vorteil der leichten Ableitbarkeit anderer Oberflächendarstellungen. So bietet unsere Speicherungsstruktur neben der Connolly-Oberfläche auch die Möglichkeit, Punkte auf der Oberfläche zu generieren. Insbesondere lassen sich Punkte auf einzelnen Patches verteilen und damit auf bestimmte Regionen der Oberfläche beschränken. Aus dem Geflecht der Kanten, die die Flächenstücke begrenzen, erhält man auf einfache Weise ein Kanten- und Flächenmodell der Molekülоberfläche, das durch allgemeine Visualisierungstools graphisch dargestellt werden kann.

Für das Dockingproblem läßt sich insbesondere die Navigationsmöglichkeit auf der Flächenstruktur gewinnbringend einsetzen, etwa zur Bestimmung von Regionen: Ausgehend von einer gegebenen Position *p* auf der Oberfläche läßt sich durch eine Umgebungsanfrage derjenige Ausschnitt aus der Oberfläche bestimmen, der diese Position umgibt, d.h. eine Zusammenhangskomponente um *p* bildet, die ein vorgegebenes Abstands- oder Ähnlichkeitskriterium erfüllt. Viele Eigenschaften, wie z.B. der Anteil bestimmter Atomtypen an den Oberflächenatomen einer Region lassen sich mit einer solchen Datenbankanfrage bestimmen.

2.2 Architektur der Anfragebearbeitung

Zur Anfragebearbeitung auf großen Datenmengen hat sich im Bereich der Geo-Datenbanksysteme das Paradigma der mehrstufigen Anfragebearbeitung bewährt. Dabei wird in der ersten Stufe, dem sog. Filterschritt versucht, eine möglichst große Anzahl von Objekten, die die Anfragebedingung nicht erfüllen, anhand eines effizient zu überprüfenden Kriteriums auszuschließen. Dadurch erhält man eine Kandidatenmenge, die eine Obermenge der eigentlichen Lösungsmenge darstellt und in den nächsten Stufen, den sog. Verfeinerungsschritten weiter untersucht wird. Dem Filterschritt kommt in der Gesamtarchitektur des Systems eine besondere Bedeutung zu, da er in hohem Maße die Gesamtlaufzeit einer Anfrage bestimmt. Er muß effizient zu berechnen sein, hohe Selektivität besitzen, darf aber keine Treffer fälschlicherweise ausschließen.

Für das Docking ist die geometrische Komplementarität der beiden Partnermoleküle an der Dockingstelle eine wichtige Voraussetzung. Dieses Kriterium läßt sich für eine grobe Vorsuche im Filterschritt heranziehen. Dazu werden auf dem Anfrageprotein mögliche Dockingstellen identifiziert und in der Datenbank wird nach Segmenten gesucht, die zu diesen eine komplementäre Form aufweisen. Die zugehörigen Proteine bilden dann die Kandidatenmenge und werden in den Verfeinerungsschritten weiter darauf untersucht, ob sie wirklich Wechselwirkungen mit dem Anfrageprotein eingehen können.

Abbildung 1 zeigt eine schematische Übersicht über die Anfragebearbeitung in unserem Dokkingsystem. Die einzelnen Schritte werden im folgenden näher erläutert:

Vorverarbeitung der Datenbankproteine. In der Vorverarbeitung werden die Moleküle für die Anforderungen von Filter- und Verfeinerungsschritten aufbereitet. Dies geschieht für alle Proteine der Datenbank im voraus, für das Anfrageprotein erst zu Beginn der Anfragebearbeitung. Die Vorverarbeitung umfaßt folgende Schritte:

- Berechnung der Proteinoberflächen.
- Segmentierung der Oberflächen, d.h. Identifizierung und Extraktion möglicher Dokkingstellen.
- Berechnung von kompakten Formbeschreibungen aus den Segmenten. Diese erlauben eine grobe Vorsuche auf den Proteinen im Filterschritt.
- Speicherung der Formbeschreibungen in der Datenbank unter Verwendung von mehrdimensionalen Indexstrukturen.

Zur Segmentierung der Proteinoberflächen wurden verschiedene alternative Verfahren entwikkelt.

- Geometrische Segmentierung mit Hilfe der konvexen Hülle eines Moleküls. Dabei werden Konkavitäten daran erkannt, daß sie von großen Facetten der konvexen Hülle über-

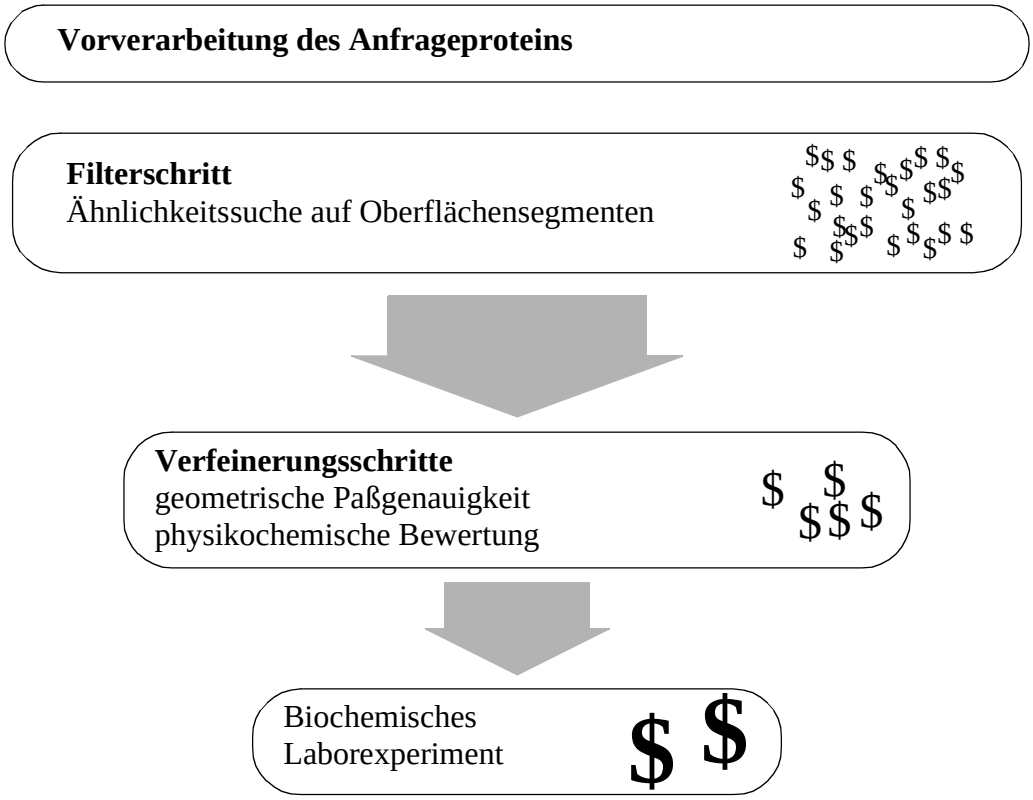

Abb. 1: Architektur der Anfragebearbeitung

spannt werden, wohingegen sich konvexe Bereiche aus den kleineren Facetten ableiten lassen [AHPS 95].

- Statistische Klassifikation und Segmentierung der Moleküloberflächen mit Methoden der Bildverarbeitung: Oberflächenpunkte mit ähnlichen Kennzahlen werden durch einen Vektorquantisierer zu Bereichen zusammengefaßt [AHPS 95].
- Clustering der Moleküloberflächen in Bereiche ähnlicher Kennzahlen durch Methoden des Knowledge Discovery. Das Clusteringverfahren CLARANS wurde für diesen Zweck so erweitert, daß es Cluster von Punkten im 3D liefert, die sowohl ähnliche Kennzahlen besitzen, als auch räumlich benachbart liegen [EKX 95a] [EKX 95b] [Wir 95].
- Segmentierung durch Regionenwachstum: Ausgehend von charakteristischen Oberflächenpunkten werden Regionen durch Erweiterung um benachbarte Punkte gebildet, bis Regionengrenzen aneinanderstoßen. Dieses Verfahren wird durch die TriEdge-Datenstruktur unterstützt [SK 95a] [SK 95b].

Vorverarbeitung des Anfrageproteins. Die Bearbeitung einer Dockinganfrage beginnt mit der Vorverarbeitung des Anfrageproteins. Wie für die Datenbankproteine bereits geschehen, wird für das Anfrageprotein die Oberfläche berechnet. Darauf werden die möglichen Dockingbereiche bestimmt, deren Formbeschreibungen den Ausgangspunkt für den Filterschritt bilden.

Filterschritt. Das Ziel dieses Schrittes ist, für das Anfrageprotein aus einer Datenbank mit Tausenden von Proteinen möglichst effizient eine kleine Menge von Oberflächenbereichen als Dockingkandidaten zu bestimmen. Dazu werden die Formbeschreibungen der Segmente des Anfrageproteins mit den Beschreibungen der Segmente in der Datenbank verglichen. Die dazu verwendeten Verfahren werden ausführlich im nächsten Abschnitt dargestellt.

Verfeinerungsschritte. Nachdem im Filterschritt Paare von möglichen Dockingbereichen bestimmt wurden, geht es in den Verfeinerungsschritten darum, die relative Lage der Partner zueinander zu bestimmen, sowie falsche Dockingvorschläge zu erkennen und zu verwerfen. So müssen etwa Vorschläge ausgemustert werden, bei denen die geometrische Paßgenauigkeit ungenügend ist, oder bei denen räumliche Überlappungen der Dockingpartner auftreten. Diese Aufgabe wird durch ein semantisches Netz gesteuert. Die eigentlichen geometrischen und physikochemischen Bewertungen werden mit Hilfe von Korrelationsanalysen sowie durch energetische Berechnungen vorgenommen. Im Unterschied zum Filterschritt, bei dem nur einzelne Oberflächenregionen betrachtet werden, arbeiten die Verfeinerungsschritte auf einer Repräsentation des gesamten Proteins.

2.3 Approximationsbasierte Formähnlichkeit von 3D-Oberflächensegmenten

Um den Filterschritt zu realisieren, mußten Verfahren erarbeitet werden, die die Ähnlichkeitssuche auf 3D-Oberflächensegmenten ermöglichen. Ein derartiges Verfahren beinhaltet zum einen ein Ähnlichkeitsmaß, das dazu geeignet ist, zu einem Anfragesegment potentielle Dokkingsegmente in der Datenbank zu identifizieren. Um die Quantifizierung der Ähnlichkeit zu ermöglichen, müssen aus den Segmenten numerische Formbeschreibungen abgeleitet werden. Die Ähnlichkeit von zwei Segmenten ergibt sich dann aus dem Vergleich dieser Formbeschreibungen. Darüber hinaus muß sich das Ähnlichkeitsmaß effizient in sehr großen Datenbanken auswerten lassen. Wir haben zwei Ansätze hierzu entwickelt und untersucht: kennzahlenbasierte und approximationsbasierte Ähnlichkeit.

Kennzahlenbasierte Ähnlichkeit. Ein Ansatz, um geometrisch ähnliche Oberflächensegmente zu bestimmen, besteht in der Verwendung von Formkennzahlen. Diese Kennzahlen beschreiben die Form eines Oberflächensegmentes in rotations- und translationsinvarianter Weise, indem sie etwa den Krümmungstyp (konkav, konvex, sattelförmig) charakterisieren. Da bei der Dockingsuche komplementäre Segmente gesucht werden, müssen sich die Kennzahlen auch leicht komplementieren lassen. Für unsere Experimente haben wir u.a. den Solid Angle (SA) ausgewählt [Ald+ 94] [EKSX 95] [EKW 96] [Wir 95], der den Öffnungswinkel der Oberfläche um einen Punkt P beschreibt. Es lassen sich auch physikochemische Kennzahlen, wie z.B. Hydrophobizität miteinbeziehen. Um mit Hilfe der Kennzahlen ein Oberflächensegment zu charakterisieren, berechnen wir zunächst für ausgewählte Punkte des Segmentes deren Kennzahlen. Danach wird eine kompakte Beschreibung der Kennzahlenverteilung eines Segments bestimmt. Ein Vergleich von zwei Segmenten erfolgt über einen Vergleich dieser Beschreibungen. Es hat sich jedoch gezeigt, daß die damit erreichbare Selektivität nicht ausreicht, um insgesamt akzeptable Laufzeiten zu erreichen.

Approximationsbasierte Ähnlichkeit. Bei der approximationsbasierten Ähnlichkeit wird eine lage- und orientierungsinvariante Formbeschreibung der Proteine aus einer lokalen Approximation der Oberfläche durch ein einfaches geometrisches Modell abgeleitet. Dieser Ansatz eignet sich zur formbasierten Suche von 3D-Oberflächensegmenten, die durch Punktmengen repräsentiert werden.

Um eine kompakte Formbeschreibung zu erhalten, werden die Segmente durch eine zweidimensionale Flächenfunktion approximiert. Ein neues, von uns entwickeltes Ähnlichkeitsmaß, das auf dieser Approximation basiert, ermöglicht eine schnelle und effektive Suche nach formähnlichen Segmenten in einer großen Datenbank. Das Verfahren soll im folgenden näher erläutert werden.

Approximation von Oberflächensegmenten. Sei $s = \{p\}$ eine Menge von 3D-Punkten, die ein Oberflächensegment eines Proteins repräsentiert. $f_{app}(x,y)$ sei das Approximationsmodell,

d.h. eine zweidimensionale multiparametrische Funktion mit dem Parametersatz $app = \langle a_i \rangle$. Das Approximationsmodell wird im voraus gewählt und sowohl auf alle Segmente in der Datenbank als auch auf die Segmente des Anfrageproteins angewandt. Die Anzahl der Parameter $\langle a_i \rangle$ von f ergibt die Dimension des Approximationsmodells. Die Approximation von s in bezug auf f wird durch die Methode der kleinsten Fehlerquadrate bestimmt, d.h. der Approximationsfehler

$$d_s^2(app) := \frac{1}{n} \sum_{p \in s} (\mathrm{f}_{app}(p_x, p_y) - p_z)^2$$

wird minimiert. Da die Segmente eine beliebige Position im 3D einnehmen können, müssen sie zur Berechnung der Approximation einer Lagenormierung unterzogen werden.

Aus numerischer Sicht sollte das Approximationsmodell so gewählt werden, daß das Minimierungsproblem, d.h. die Berechnung der Approximation, effizient gelöst werden kann. Dies ist für solche Modelle der Fall, die als Linearkombination der Approximationsparameter a_i mit beliebigen zweidimensionalen Funktionen $f_i\,(x,y)$ dargestellt werden können. Beispiele hierfür sind Paraboloide wie $a_1 \cdot x^2 + a_2 \cdot y^2$ für Grad zwei, oder trigonometrische Polynome wie $a_1 \cdot cos(x) + a_2 \cdot sin(x) + a_3 \cdot cos(y) + a_4 \cdot sin(y)$. Die allgemeine Formel für derartige Modelle ist:

$$F_{app}(x,y) = \sum_{i\,=\,1 \ldots k} app_i \cdot \mathrm{f}_i(x, y)$$

Ähnlichkeit von Segmenten. Mit Hilfe der berechneten Approximationen kann die Ähnlichkeit auf Segmenten definiert werden. Eine einfache Möglichkeit zur Definition der Ähnlichkeit zweier Segmente s und q wäre der Abstand der Vektoren, die aus den jeweiligen Approximationsparametern gebildet werden. Dabei tritt jedoch das Problem auf, daß es für ein Segment viele Approximationsparametersätze geben kann, deren Fehler nah am Minimum liegt, die sich in ihren Werten jedoch erheblich unterscheiden. Der alleinige Vergleich der Approximationsparameter liefert in solchen Fällen keine befriedigenden Ergebnisse. Um den Abstand der beiden Segmente s und q zu quantifizieren, wird daher jedes Segment mit der Approximation des anderen Segmentes verglichen. Das Abstandsmaß $\mathrm{d}_{app}(s, q)$ ergibt sich dann als Summe der wechselseitigen Approximationsfehler.

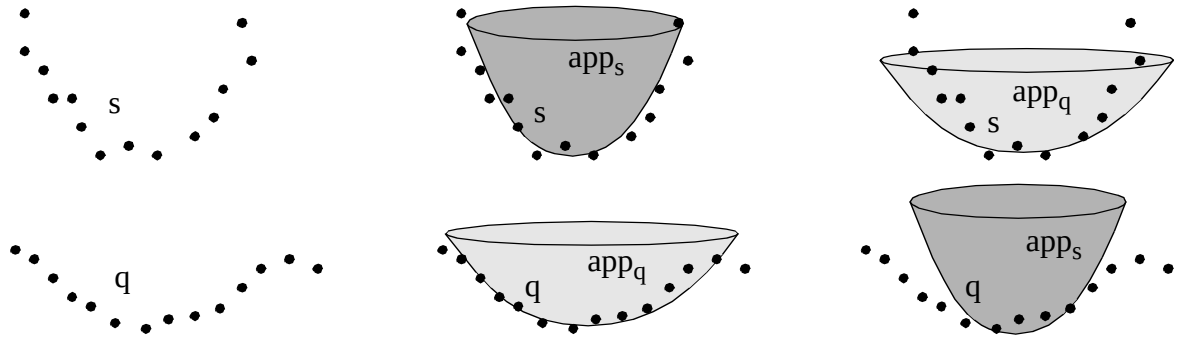

Abb. 2: Ähnlichkeitsquantifizierung durch gegenseitigen Approximationsabstand

Ein weiteres wichtiges Kriterium für die Auswahl von möglichen Partnersegmenten für ein Anfragesegment ist deren Größe. Diese läßt sich z.B. durch die Ausdehnung eines Segments in den Hauptachsen festlegen. Der Größenabstand $\mathrm{d}_{ext}(s, q)$ ergibt sich dann als euklidischer Abstand der (dreidimensionalen) Ausdehnungsvektoren.

Um sowohl Form als auch Größe der Segmente berücksichtigen zu können, ist eine Kombination aus beiden Maßen sinnvoll. Das Ähnlichkeitsmaß d_{shape} (s, q) für approximationsbasierte Formähnlichkeit lautet daher (mit zwei Gewichtungsfaktoren u_{app} und u_{ext}):

$$d_{shape}(s, q) = \sqrt{u_{app} d^2_{app}(s, q) + u_{ext} d^2_{ext}(s, q)}$$

Effiziente Anfragebearbeitung. Zur effizienten Anfragebearbeitung auf multidimensionalen Datenobjekten ist die Verwendung von Indexstrukturen wie mehrdimensionale Punkt- oder Raumzugriffsstrukturen unerläßlich. In [KSS 97] wird gezeigt, daß sich auch d_{shape} effizient auf einer Indexstruktur auswerten läßt. Die Suche nach allen Segmenten, deren Abstand bezüglich d_{shape} kleiner als ein bestimmter Wert ist, ergibt eine Ellipsoidanfrage, d.h. eine Bereichsanfrage mit einem ellipsoidförmigen Anfragebereich. Wir haben Methoden entwikkelt, wie sich derartige Anfragen auf einer Indexstruktur auswerten lassen, die ein Directory mit Hilfe von achsenparallelen Rechtecken verwaltet.

2.4 Flexible Ellipsoidanfragen

Bei der Entwicklung der effizienten Bearbeitungstechnik für approximationsbasierte Ähnlichkeitsanfragen ergab sich die Ellipsoidanfrage als eine Basisoperation im Filterschritt. Weitergehende Untersuchungen zeigten, daß dieser neue Anfragetyp eine grundlegende und weitreichende Bedeutung für die Ähnlichkeitssuche in Multimedia- und Geodatenbanksystemen darstellt. Vor allem für kennzahl- und histogrammbasierte Ähnlichkeitsmodelle lassen sich Ellipsoidanfragen nutzbringend einsetzen. Die Ellipsoide entsprechen quadratischen Formen, die durch Ähnlichkeitsmatrizen spezifiziert sind. Im Gegensatz zum klassischen Euklidischen Abstand lassen sich diese allgemeineren Distanzfunktionen besser an die Anforderungen bestimmter Anwendungen anpassen. Unsere neuen Algorithmen unterstützen insbesondere die Anpassung der Ähnlichkeitsmatrix zur Anfragezeit als eine neue und mächtige Flexibilität für die Ähnlichkeitssuche. Neben der approximationsbasierten Ähnlichkeitssuche [KSS 97] [KS 98] wurde dieses Modell bereits auf Farbhistogramme in Bilddatenbanken [SK 97] sowie auf ein einfaches Molekülmodell in 3D-Proteindatenbanken [Sei 97] erfolgreich angewendet.

3 Weitere Aspekte

Nach der eingehenden Darstellung des im Projekt erzielten Ergebnisses im vorhergehenden Abschnitt werden hier nun die weiteren Aspekte nach Nr. 5b.2.2 BNBest dargelegt: Der voraussichtliche Nutzen des Ergebnisses, der Fortschritt bei anderen Stellen, sowie die bereits erfolgten bzw. geplanten Veröffentlichungen des Ergebnisses.

3.1 Nutzen des Projektergebnisses

Das Verbundprojekt BIOWEPRO war als sehr ehrgeiziges Vorhaben zwischen Anwendungsorientierung und Grundlagenforschung angesiedelt. Der voraussichtliche Nutzen des Ergebnisses liegt im Zeit- und Kostensparpotential, das die 1:n-Dockingsuche für die Produktenwicklungen in der chemischen und pharmazeutischen Industrie bedeutet. Aufgrund der besonderen Schwierigkeiten der 1:n-Dockingvorhersage bedürfen die im Projekt entwikkelten Techniken jedoch noch weiterer Untersuchung auf einer breiten experimentellen Basis. Insofern ist die direkte industrielle Verwertbarkeit des Ergebnisses noch nicht gegeben. Auf jeden Fall sollten jedoch die Erfahrungen aus dem Projekt für weitere Entwicklungen genutzt

werden. Insbesondere das Konzept der approximationsbasierten Ähnlichkeit bietet eine vielversprechende Grundlage für brauchbare 1:n-Dockingverfahren.

3.2 Fortschritt bei anderen Stellen

Während der Durchführung des Vorhabens wurden weder dem Zuwendungsempfänger noch den anderen Projektgruppen im Verbundprojekt BIOWEPRO Arbeiten auf dem Gebiet des 1:n-Docking bei anderen Stellen bekannt. Zwar wurde die Aufgabenstellung des 1:1-Docking in mehreren Gruppen bearbeitet [FNNW 93] [KLR 95] [RKL 95], doch sind uns keine Arbeiten bekannt, die sich bislang mit dem weitergehenden Problem der 1:n-Dockingsuche in großen Proteindatenbanken befaßt haben.

In Bezug auf die Datenbanktechniken zur Ähnlichkeitssuche wurden aktuelle Entwicklungen aufgegriffen. Insbesondere neue Algorithmen zur mehrstufigen Bearbeitung von Ähnlichkeitsanfragen wurden analysiert [FRM 94] [HS 95] [Kor+ 96] und erfolgreich für unsere neuen Ähnlichkeitsmodelle eingesetzt [KSS 97]. Im Fall der *k*-nächsten-Nachbarn-Suche konnte sogar eine wesentliche Verbesserung erreicht werden [SK 98].

3.3 Veröffentlichung des Ergebnisses

Schon während der Projektlaufzeit wurden immer wieder die aktuellen Ergebnisse unserer Untersuchungen und Entwicklungen sowohl auf nationalen als auch auf internationalen Fachkongressen veröffentlicht und so den fachlich interessierten Stellen zugänglich gemacht. Die einzelnen Publikationen lassen sich den folgenden Bereichen zuordnen: Molekülоberflächen in Proteindatenbanken [KSS 95] [SK 95a] [SK 95b], Architekturen und Algorithmen zur Dokkingsuche und geometrischen Ähnlichkeitssuche [Ald+ 94] [EKSX 95] [KSS 96] [KSS 97] [Sch+ 95], sowie Wissensextraktion in Geodatenbanksystemen, angewandt auf Proteindatenbanken [EKSX 96] [EKX 95a] [EKX 95b] [EKW 96]. Weiterführende Arbeiten, die durch das Projekt inspiriert wurden, sind geplant oder bereits erfolgt [AKS 98] [KS 98] [Sei 97] [SK 97] [SK 98]. Zusätzlich zu den genannten Veröffentlichungen wurden eine Reihe von Diplomarbeiten [San 96] [Sch 94] [Som 94] [Wir 95] im Rahmen des Projektes angeregt und betreut.

4 Literaturangaben

Im folgenden Verzeichnis ist die im Projekt verwendete Literatur aus den Bereichen Molekulare Bioinformatik und Datenbanktechniken sowie die im Projekt erfolgten und geplanten Veröffentlichungen und die betreuten Diplomarbeiten aufgeführt.

4.1 Verwendete Literatur zur Molekularen Bioinformatik

[AHPS 95] Ackermann F., Herrmann G., Posch S., Sagerer G.: *‘3D-Segmentierungstechniken und vektorwertige Bewertungsfunktionen für symbolisches Protein-Protein-Docking’*, Proc. 3. GI-Fachtagung ‘Bioinformatik - Informatik in den Biowissenschaften’, Braunschweig, 09.-11.10.95, in: *Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism.* GBF Monographs, Vol. 18, 1995, pp. 105-124.

[Ber+ 77] Bernstein F. C., Koetzle T. F., Williams G. J., Meyer E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanovichi T., Tasumi M.: *‘The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures’*, Journal of Molecular Biology, Vol. 112, 1977, pp. 535-542.

[BMH 92] Badel A., Mornon J. P., Hazout S.: *'Searching for Geometric Molecular Shape Complementarity using Bidimensional Surface Profiles'*, Journal of Molecular Graphics, Vol. 10, 1992, pp. 205-211.

[Con 83] Connolly M. L.: *'Solvent-Accessible Surfaces of Proteins and Nucleic Acids'*, Science, Vol. 221, 1983, pp. 709-713.

[Con 86] Connolly M. L.: *'Shape Complementarity at the Hemoglobin $\alpha_1\beta_1$ Subunit Interface'*, Biopolymers, Vol. 25, 1986, pp. 1229-1247.

[FNNW 93] Fischer D., Norel R., Nussinov R., Wolfson H. J.: *'3-D Docking of Protein Molecules'*, Proc. 4th Annual Symp. on Combinatorial Pattern Matching (CPM '93), Padova, Italy, in: Lecture Notes in Computer Science, Vol. 684, Springer, 1993, pp. 20-34.

[HO 94] Halperin D., Overmars M. H.: *'Spheres, Molecules, and Hidden Surface Removal'*, Proc. 10th ACM Symp. Computational Geometry, 1994, pp. 113-122.

[Kat+ 92] Katchalski-Katzir E., Shariv I., Eisenstein M., Friesem A. A., Aflalo C., Vakser I. A.: *'Molecular Surface Recognition: Determination of Geometric Fit between Proteins and their Ligands by Correlation Techniques'*, Proc. National Academy of Science USA, Vol. 89, 1992, pp. 2195-2199.

[KLR 95] Klebe G., Lengauer T., Rarey M.: *'Computational Methods for Molecular Docking'*, Tutorial Notes, 3rd Int. Conference on Intelligent Systems for Molecular Biology (ISMB-95), Cambridge, UK, July 1995.

[SBK 92] Shoichet B. K., Bodian D. L., Kuntz I. D.: *'Molecular Docking Using Shape Descriptors'*, Journal of Computational Chemistry, Vol. 13, No. 3, 1992, pp. 380-397.

[RKL 95] Rarey M., Kramer B., Lengauer T.: *'Time-Efficient Docking of Flexible Ligands into Active Sites of Proteins'*, Proc. 3rd Int. Conference on Intelligent Systems for Molecular Biology (ISMB–95), Cambridge, UK, AAAI Press, 1995, pp. 300-308.

[VBW 94] Varshney A., Brooks F. P., Wright W. V.: *'Computing Smooth Molecular Surfaces'*, IEEE Computer Graphics & Applications, Vol. 14, No. 5, 1994, pp. 19-25.

4.2 Verwendete Literatur zu Datenbanktechniken

[AFS 93] Agrawal R., Faloutsos C., Swami A.: *'Efficient Similarity Search in Sequence Databases'*, Proc. 4th. Int. Conf. on Foundations of Data Organization and Algorithms (FODO '93), Evanston, ILL, in: Lecture Notes in Computer Science, Vol. 730, Springer, 1993, pp. 69-84.

[BKSS 90] Beckmann N., Kriegel H.-P., Schneider R., Seeger B.: *'The R*-tree: An Efficient and Robust Access Method for Points and Rectangles'*, Proc. ACM SIGMOD Int. Conf. on Management of Data, Atlantic City, NJ, 1990, pp. 322-331.

[BHKS 93] Brinkhoff T., Horn H., Kriegel H.-P., Schneider R.: *'Eine Speicher- und Zugriffsarchitektur für effiziente Anfragebearbeitung in Geo-Datenbanksystemen'*, Proc. GI-Fachtagung Datenbanksysteme in Büro, Technik und Wissenschaft, Braunschweig, in: Informatik aktuell, Springer, 1993, S. 356-374.

[BK 94] Brinkhoff T., Kriegel H.-P.: *'The Impact of Global Clustering on Spatial Database Systems'*, Proc. 20th Int. Conf. on Very Large Data Bases, Santiago, Chile, 1994, pp. 168-179.

[BKK 96] Berchtold S., Keim D., Kriegel H.-P.: *'The X-tree: An Index Structure for High-Dimensional Data'*, Proc. 22nd Conf. on Very Large Data Bases, Mumbai, India, 1996, pp. 28-39.

[BS 97] Barequet G., Sharir M.: *'Partial Surface and Volume Matching in Three Dimensions'*, IEEE Trans. on Pattern Analysis and Machine Intelligence (T-PAMI), Vol. 19, No. 9, 1997, pp. 929-948.

[FRM 94] Faloutsos C., Ranganathan M., Manolopoulos Y.: *'Fast Subsequence Matching in Time-Series Databases'*, Proc. ACM SIGMOD Int. Conf. on Management of Data, Minneapolis, MN, 1994, pp. 419-429.

[HS 95] Hjaltason G. R., Samet H.: *'Ranking in Spatial Databases'*, Proc. 4th Int. Symp. on Large Spatial Databases, Portland, ME, 1995, pp. 83-95.

[Kor+ 96] Korn F., Sidiropoulos N., Faloutsos C., Siegel E., Protopapas Z.: *'Fast Nearest Neighbor Search in Medical Image Databases'*, Proc. 22nd VLDB Conference, Mumbai (Bombay), India, 1996, pp. 215-226.

[KSB 93] Kriegel H.-P., Schneider R., Brinkhoff T.: *'Potentials for Improving Query Processing in Spatial Database Systems'*, Invited talk, Proc. 9emes Journées Bases de Données Avancées (9th Conference on Advanced Databases), Toulouse, France, 1993.

[OHMS 92] Orenstein J., Haradhvala S., Margulies B., Sakahara D.: *'Query Processing in the ObjectStore Database System'*, Proc. ACM SIGMOD 1992, pp. 403-412.

[Ore 86] Orenstein J: *'Spatial Query Processing in an Object-Oriented Database System'*, Proc. ACM SIGMOD Int. Conf. on Management of Data, 1986, pp. 326-336.

[SKSH 89] Schneider R., Kriegel H.-P., Seeger B., Heep S.: *'Geometry-based Similarity Retrieval of Rotational Parts'*, Proc. Int. Conf. on Data and Knowledge Systems for Manufacturing and Engineering, Gaithersburg, ML, 1989, pp. 150-160.

4.3 Veröffentlichungen der BIOWEPRO-Projektgruppe

[AKS 98] Ankerst M., Kriegel H.-P., Seidl T.: *'Pixel-based Shape Similarity Search in Large Image Databases'*, eingereicht zur Veröffentlichung

[Ald+ 94] Aldinger K., Ester M., Förstner G., Kriegel H.-P., Seidl T.: *'Datenbankunterstützung für das Protein-Protein-Docking: ein effizienter und robuster Feature-Index'*, Proc. 'Bioinformatik – Computereinsatz in den Biowissenschaften', 2. GI-Fachtagung 'Informatik in den Biowissenschaften', Jena, 05.-07.09.1994, S. 41-52.

[EKSX 95] Ester M., Kriegel H.-P., Seidl T., Xu X.: *'Formbasierte Suche nach komplementären 3D-Oberflächen in einer Protein-Datenbank'*, Proc. GI-Fachtagung 'Datenbanken in Büro, Technik und Wissenschaft' (BTW'95), Reihe Informatik aktuell, Springer, 1995, S. 373-382.

[EKSX 96] Ester M., Kriegel H.-P., Sander J., Xu X.: *'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'*, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96).

[EKX 95a] Ester M., Kriegel H.-P., Xu X.: *'Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification'*, Proc. 4th Int. Symp. on Large Spatial Databases (SSD'95), Portland, Maine, USA, Lecture Notes in Computer Science, Vol. 951, Springer, 1995, pp. 67-82.

[EKX 95b] Ester M., Kriegel H.-P., Xu X.: *'A Database Interface for Clustering in Large Spatial Databases'*, Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD–95), Montreal, Canada, AAAI Press, 1995.

[EKW 96] Ester M, Kriegel H.-P., Wirth S.: *'Feature Based Classification of Protein Docking Sites: An Algorithm for Large Databases and Experimental Results'*, Proc. German Conf. on Bioinformatics (GCB'96), 1996.

[KS 98] Kriegel H.-P., Seidl T.: *'Approximation-based Similarity Search for 3-D Surface Segments'*, eingereicht zur Veröffentlichung.

[KSS 95] Kriegel H.-P., Schmidt T., Seidl T.: *'Molekülоberflächen in einer objektorientierten Proteindatenbank'*, Proc. 3. GI-Fachtagung 'Bioinformatik—Informatik in den Biowissenschaften', Braunschweig, 09.-11.10.1995, erweiterte Zusammenfassung.

[KSS 96] Kriegel H.-P., Schmidt T., Seidl T.: *'Können Proteine im Computer docken?'*, in: Einsichten, Ludwig-Maximilians-Universität München, ISSN 0941-3642, Vol. 2, 1996, pp. 26-29.

[KSS 97] Kriegel H.-P., Schmidt T., Seidl T.: *'3D Similarity Search by Shape Approximation'*, Proc. Fifth Int. Symp. on Large Spatial Databases (SSD'97), Berlin, Germany, Lecture Notes in Computer Science, Vol. 1262, 1997, pp. 11-28.

[Sch+ 95] Schomburg D., Jakob U., Meyer M., Wilson P., Sagerer G., Ackermann F., Herrmann G., Posch S., Soumpasis M., Grimm G., Ihmels B., Strahm M., Kriegel H.-P., Seidl T., Schmidt T., Ester M., Xu X.: *'BIOWEPRO – Biomolekulare Wechselwirkungen von Proteinen'*, in BMBF (Hrsg.): 'Molekulare Bioinformatik und Evolutionäre Algorithmen', Tagungsband BMBF-Statusseminar Bioinformatik, Braunschweig, 7.-8. Nov. 1995, S. 125-153.

[Sei 97] Seidl T.: *'Adaptable Similarity Search in 3-D Spatial Database Systems'*, Dissertation, Fakultät für Mathematik und Informatik, Universität München, 1997.

[SK 95a] Seidl T., Kriegel H.-P.: *'Solvent Accessible Surface Representation in a Database System for Protein Docking'*, Proc. 3rd Int. Conf. on Intelligent Systems for Molecular Biology (ISMB–95), Cambridge, UK, AAAI Press, 1995, pp. 350-358.

[SK 95b] Seidl T., Kriegel H.-P.: *'A 3D Molecular Surface Representation Supporting Neighborhood Queries'*, Proc. 4th Int. Symp. on Large Spatial Databases (SSD'95), Portland, Maine, USA, Lecture Notes in Computer Science, Vol. 951, Springer, 1995, pp. 240-258.

[SK 97] Seidl T., Kriegel H.-P.: *'Efficient User-Adaptable Similarity Search in Large Multimedia Databases'*, Proc. 23rd Int. Conf. on Very Large Databases (VLDB'97), Athens, Greece, 1997, pp. 506-515.

[SK 98] Seidl T., Kriegel H.-P.: *'Optimal Multi-Step k-Nearest Neighbor Search'*, submitted for publication.

4.4 Im Projekt BIOWEPRO betreute Diplomarbeiten

[San 96] Sander J.: *'Dichtebasiertes Clustering zum Knowledge-Discovery in Geo-Datenbanksystemen'*, Diplomarbeit, Institut für Informatik, Universität München, 1996.

[Sch 94] Schmidt T.: *'Berechnung von Proteinoberflächen mit Hilfe räumlicher Indexstrukturen'*, Diplomarbeit, Institut für Informatik, Technische Universität München, 1994.

[Som 94] Sommer U.: *'Objektorientierte Modellierung einer Proteindatenbank unter Berücksichtigung der 3D-Struktur'*, Diplomarbeit, Institut für Informatik, Technische Universität München, 1994.

[Wir 95] Wirth S.: *'Algorithmen zur Konstruktion von Assoziationsregeln in Geo-Datenbanken und deren Anwendung in einer Proteindatenbank'*, Diplomarbeit, Institut für Informatik, Universität München, 1995.

Prof. Dr. Hans-Peter Kriegel