GI-Fachtagung “Datenbanken in Büro, Technik und Wissenschaft” (BTW ‘95), *Informatik Aktuell*, Springer, 1995.

Formbasierte Suche nach komplementären 3D-Oberflächen in einer Protein-Datenbank[1]

Martin Ester, Hans-Peter Kriegel, Thomas Seidl, Xiaowei Xu

Institut für Informatik, Universität München

Leopoldstr. 11B, D-80802 München

{ ester | kriegel | seidl | xu }@informatik.uni-muenchen.de

Zusammenfassung:
Die Komplementarität der 3D-Oberflächen von Proteinen ist neben den physikochemischen Eigenschaften ein entscheidendes Kriterium dafür, ob und an welchen Stellen zwei Proteine miteinander wechselwirken, d.h. docken, können. Anders als in Geo-Datenbanksystemen, die Anfragen nach Objekten mit einer gegebenen räumlichen Lage und Ausdehnung unterstützen, werden deshalb beim Protein-Docking Objekte aufgrund ihrer Form gesucht. Wir beginnen mit einer Darstellung der Anforderungen dieser neuen Anwendung an Datenbanksysteme. Zur effizienten Anfragebearbeitung übernehmen wir die in Geo-Datenbanksystemen bewährte Technik der mehrstufigen Anfragebearbeitung, die den Kreis der potentiellen Dockingkandidaten sehr schnell einengt. Der Filterschritt benutzt eine Beschreibung der geometrischen Oberflächencharakteristik mit Hilfe rotations- und translationsunabhängiger Kennzahlen. Ein mehrdimensionaler Join liefert Dockingpartner, die Oberflächenpunkte mit komplementären Kennzahlwerten besitzen. Im Verfeinerungsschritt werden selektierte Oberflächenpunkte mit ihren Nachbarn zu 3D-Regionen erweitert und deren Komplementarität gemessen. Die mehrstufige Anfragebearbeitung wird in ein Docking-System integriert, das auf einem kommerziellen objektorientierten Datenbanksystem basiert.

Schlüsselwörter: Anfragebearbeitung in Geo-Datenbanksystemen, Datenbanken in der Biologie, Dockingsuche in Protein-Datenbanken, Ähnlichteilsuche, 3D-Formbeschreibungen.

1 Einleitung

Die Funktion eines Proteins besteht in der Interaktion mit anderen Biomolekülen, etwa wieder Proteinen, DNA oder auch kleineren Partnern. Da die Interaktionen eine Anlagerung der betroffenen Partner aneinander darstellen, bezeichnet man sie als *Docking*. Wir entwickeln ein Protein-Docking-Datenbanksystem, das für ein vorgegebenes Anfrageprotein mögliche Interaktionspartner in einer Datenbank von Proteinen finden soll. Neben den physikochemischen Eigenschaften der Moleküle spielt die Geometrie der Moleküloberfläche an der Interaktionsstelle eine zentrale Rolle beim Docking. Das Datenbanksystem muß deshalb Anfragen nach Proteinen mit ähnlichen bzw. komplementären Oberflächen beantworten.

Geo-Datenbanksysteme dienen der Verwaltung von räumlich ausgedehnten Objekten in Anwendungen wie CAD-Datenbanken, geographischen Informationssystemen oder Bild-Datenbanken. Typische Anfragen in Geo-Datenbanksystemen suchen Objekte mit einer definierten räumlichen Lage, z.B. liefert eine Region-Query alle Objekte, die eine gegebene Anfrageregion schneiden und ein Spatial Join [BKSS 94] alle Paare von Objekten, die in einer gegebenen räumlichen Beziehung zueinander stehen. Zur effizienten Anfragebearbeitung in Geo-Datenbanksystemen haben wir neben räumlichen Zugriffsstrukturen Methoden der Approximation, der Dekomposition und der Clusterung [BK 94] entwickelt. Zur Beantwortung von Anfragen

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

nach Objekten mit gleicher oder ähnlicher Form unabhängig von der räumlichen Lage, wie in unserer Anwendung erforderlich, sind bisherige Algorithmen der Anfragebearbeitung in Geo-Datenbanksystemen jedoch nicht geeignet.

Es gibt in der Literatur einige Vorschläge, Geo-Datenbanksysteme zur Unterstützung von Anfragen nach der Form von Objekten zu erweitern. Zwei Ansätze werden dazu verfolgt:

- Normierung der Lage im geometrischen Raum
  [Jag 91] und [MG 93] beschreiben die Form der Objekte durch eine geordnete Menge von Punkten bzw. Rechtecken, deren Lage im Raum normiert wird. Die Punkte bzw. Rechtecke werden in Punkte eines höherdimensionalen Raums transformiert und mit Hilfe einer der bekannten Punktzugriffsstrukturen abgespeichert. Die Suche nach einem Objekt mit gegebener Form wird dann als Region-Query formuliert. Beide Artikel betrachten 2D-Objekte. [SKSH 89] behandelt die Ähnlichteilsuche für rotationssymmetrische CAD-Bauteile. Die 3D-Bauteile werden durch ihre zweidimensionale Kontur repräsentiert. Die Lage der Konturen im Raum wird geeignet normiert und für die Suche genutzt.
- Abstraktion von der Lage mit Hilfe von geometrischen Kennzahlen
  Die Form eines Objekts wird nicht direkt durch seine Geometrie, sondern mit Hilfe von Kennzahlen beschrieben, die die Charakteristik der Geometrie in lageunabhängiger Weise beschreiben. Solche Kennzahlen können z.B. mit Hilfe einer diskreten Fourier-Transformation gewonnen werden, wie sie in [AFS 93] und [FRM 94] für eine Datenbank von Zeitreihen reeller Zahlen eingesetzt wird.

Alle obigen Verfahren arbeiten jedoch im 2D und sind somit für dreidimensionale Proteinoberflächen nicht direkt geeignet. Unser Lösungsansatz basiert auf der in Geo-Datenbanksystemen üblichen Architektur einer mehrstufigen Anfragebearbeitung (vgl. [BHKS 93], [KSB 93]). Im Filterschritt verwenden wir, wie auch in der Molekularbiologie vorgeschlagen, geometrische Kennzahlen zur Abstraktion von der Lage. Im Verfeinerungsschritt normieren wir die Lage der Objekte, um die 3D-Geometrie von Dockingkandidaten miteinander vergleichen zu können.

Der Artikel gliedert sich wie folgt: Kapitel 2 beschreibt die wichtigsten Anforderungen an Datenbanksysteme, die die Anwendung des Protein-Protein-Docking stellt. In Kapitel 3 wird die Architektur unserer Anfragebearbeitung vorgestellt. Definition, Berechnung und Komplementierung der Oberflächenkennzahlen werden in Kapitel 4 dargestellt. In Kapitel 5 werden der Filterschritt und der Verfeinerungsschritt der Anfragebearbeitung behandelt. Kapitel 6 faßt die Ergebnisse zusammen und gibt einen Ausblick.

2 Anforderungen des Protein-Protein-Docking an Datenbanksysteme

Die Vorhersage von Dockingvorgängen zwischen zwei Molekülen ist ein aktueller Forschungszweig in der Molekularbiologie, in dem neuartige Anforderungen an Datenbanksysteme auftreten. Unsere Aufgabenstellung im Projekt BIOWEPRO (Biomolekulare Wechselwirkungen von Proteinen) [Ald 94] besteht darin, für ein vorgegebenes Anfrageprotein mögliche Interaktionspartner in einer Datenbank von Proteinen zu finden. Zusätzlich soll die relative Position eines Antwortmoleküls zum Anfrageprotein angegeben werden, deren sechs Freiheitsgrade sich mit drei Translations- und drei Rotationsparametern beschreiben lassen. Ein Tupel (Protein, Protein, relative Position) bezeichnen wir als *Konstellation*. Alle Autoren stimmen darin überein, daß beim Protein-Protein-Docking neben den physikochemischen Eigenschaften der Moleküle die Geometrie der Molekülоberfläche an der Interaktionsstelle eine große Rolle spielt.

Als Datengrundlage stehen die kristallographisch ermittelten Atomkoordinaten von Proteinen und Proteinkomplexen in der Brookhaven Protein Data Bank (PDB) zur Verfügung [Ber 77]. Die PDB enthält derzeit knapp 3.000 Proteine, Enzyme und Viren [PDB 94] und wächst ständig. Die PDB ist eine Sammlung von Files, in denen jeweils die Daten eines Proteins als Text enthalten sind. Ein Datenbankmanagementsystem im Sinne von [Ull 88] ist nicht vorhanden.

Ein *Protein* besteht aus mehreren Ketten, die sich wiederum aus Aminosäuren und diese aus (insgesamt mehreren hundert bis tausend) Atomen zusammensetzen. Die Atome besitzen je nach Typ verschiedene Radien, Ladungen etc. und eine Position im 3D-Raum, gegeben durch die Koordinaten ihres Zentrums. Mit Hilfe der Eigenschaften der Atome lassen sich die physikochemischen Eigenschaften des Proteins an jedem 3D-Punkt berechnen. Proteine sind also sowohl sehr große als auch sehr komplex strukturierte Objekte.

Die *Oberfläche* eines Proteins wird als diejenige Fläche definiert, die für eine Probenkugel mit gegebenem Radius zugänglich ist [Ric 77]. Sie setzt sich aus konkaven, sattelförmigen und konvexen Flächenstücken zusammen (siehe das Beispiel in Abbildung 1).

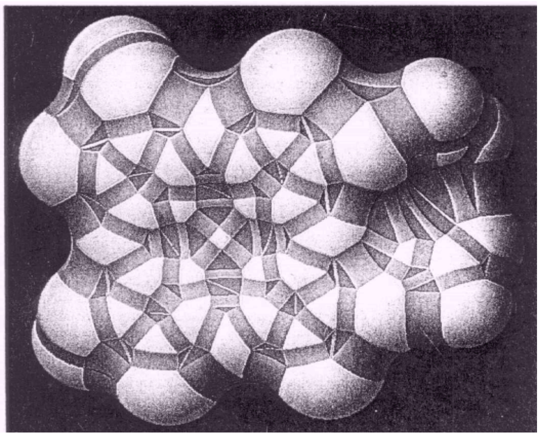

Abb. 1: Oberfläche einer Teilgruppe von Hämoglobin [Con 83].

Diese analytische Repräsentation ist z.B. für die Berechnung des Normalenvektors eines Oberflächenpunkts geeignet. Durch eine möglichst gleichmäßige Verteilung von Punkten erhält man eine gepunktete Repräsentation der Oberfläche, die z.B. für die Berechnung von geometrischen Oberflächeneigenschaften benötigt wird. Die Anwendung erfordert also eine Mehrfachrepräsentation der Proteinoberflächen in der Datenbank.

Auf den oben skizzierten Objekten der Datenbank arbeiten Dockingalgorithmen, deren zentrales Problem die enorme Größe des Suchraums aller möglichen Konstellationen ist. Die PDB enthält derzeit ca. 2500 Proteine, die in der von uns gewählten Auflösung durchschnittlich 2000 Oberflächenpunkte besitzen. Das ergibt ca. $2500 * 2000 * 2000 = 10^{10}$ verschiedene Konstellationen, die für die Bearbeitung einer Anfrage relevant sind. Die bisherigen Verfahren beschränken sich alle auf das 1:1-Docking, d.h. auf die Interaktion eines Paares von Proteinen. Wir untersuchen daher im folgenden (Übersicht siehe Abbildung 2) Algorithmen aus der Literatur auf ihre Anwendbarkeit für das 1:n-Docking, d.h. die Interaktion eines Proteins mit allen Proteinen der Datenbank.

Referenz	Verfahren	Laufzeit	Anzahl Beispiele, Erfolg (+ / –)
[Kat 92]	gitterbasierte Suche, diskrete FFT	∅ 7,5 Std.	3+, 4–
[PD 93]	Simulated Annealing	∅ 30 Std.	3+
[Con 86b]	Matchen von Punkten (“knobs and holes”)	17 Std.	1+
[BMH 92]	Matchen von Profilen; nur 2D, kein 3D	0,3 Std.	1 ?
[HT 94]	Matchen von Schnitten	3 - 24 Std.	3+, 1–
[FNNW 93]	Matchen von Dreiecken im 3D	0,5-0,75 Std.	1+, 1–

Abb. 2: Vergleich von ausgewählten Dockingverfahren

Bei den Verfahren in [Kat 92] und [PD 93] werden für zu untersuchende Proteinpaare alle Schritte der Dockingsuche jedesmal neu durchlaufen. Es ist keine Vorverarbeitung möglich, so daß sich die Ansätze nicht für das 1:n-Docking eignen. Solche Verfahren eignen sich also besser für Verfeinerungsschritte als für die erste Suche nach Dockingpartnern und -stellen. Wegen der immensen Größe des Suchraums eignen sich für den ersten Filterschritt einer Datenbanksuche besser solche Methoden, die eine Vorbearbeitung der Moleküle durchführen. Die Ermittlung charakteristischer Formbeschreibungen sowie eine translations- und rotationsinvariante Darstellung ermöglichen eine Indexunterstützung für die Suche.

In der Literatur finden sich dazu einige Anregungen. [Con 86b] berechnet auf der Proteinoberfläche "knobs and holes" als Punkte mit lokal extremen Werten für "Solid Angle" (vgl. Kapitel 4). Diese Punkte charakterisieren Aus- und Einbuchtungen auf der Oberfläche und eignen sich für die paarweise Gegenüberstellung. Auf diese Weise lassen sich erfolgversprechende Dockingkonstellationen ohne vollständiges Durchlaufen eines Rasters ermitteln. [BMH 92] und [HT 94] beschreiben die Formen von Molekülen durch 2D-Schnittkonturen mit Hilfe von Polygonen, auf denen in der Anfragebearbeitung ähnliche Ausschnitte ermittelt werden. Die Verallgemeinerung von 2D auf 3D ist dabei noch nicht zufriedenstellend gelöst. [FNNW 93] arbeitet mit Distanzmatrizen und "geometrischem Hashing" und ist leider nicht ausreichend robust.

Ein allgemeiner Algorithmus zur Lösung des Protein-Dockingproblems ist derzeit noch nicht absehbar. Die "Korrektheit" von Algorithmen wird in der Regel nur an wenigen bekannten Beispielkomplexen überprüft. Diese Prüfung ist ein grundlegendes Problem, da die Korrektheit einer Vorhersage für unbekannte Dockingpartner nur im biochemischen Laborexperiment nachgewiesen werden kann und sehr zeitaufwendig ist.

Zusammenfassend läßt sich feststellen, daß das Protein-Protein-Docking eine große Herausforderung für Datenbanksysteme darstellt. Folgende *Anforderungen* sind zu erfüllen:

- Verwaltung komplexer Objekte
- Verwaltung von 3D-Freiformflächen
- effiziente 3D-Ähnlichkeitssuche.

Im Rahmen dieses Artikels konzentrieren wir uns auf Ansätze zur Erfüllung der letzten Anforderung.

3 Architektur der Anfragebearbeitung

Die enorme Größe des Konstellationsraums, die in Kapitel 2 skizziert wurde, legt eine Anfragebearbeitung in mehreren Schritten nahe, um akzeptable Laufzeiten zu erreichen. Wegen der anfänglich sehr großen Menge potentieller Konstellationen wird in den ersten Schritten auf ihre genaue Untersuchung zu Gunsten geringerer Filterkosten pro Konstellation verzichtet. Wir arbeiten nicht auf der komplexen 3D-Oberfläche, sondern auf einer Abstraktion mit Hilfe geometrischer Oberflächenkennzahlen. Wegen der immer kleiner werdenden Menge von potentiellen Dockingstellen dürfen die späteren Schritte mit höheren Kosten pro Dockingkandidat verbunden sein und können deshalb auf der 3D-Oberfläche selbst arbeiten.

Da bisher kein Algorithmus existiert, der mit Sicherheit feststellt, ob eine Konstellation dockt, müssen die vom Dockingsystem gelieferten Resultate letztendlich vom Molekularbiologen im Experiment überprüft werden. Im Unterschied zu Geo-Datenbanksystemen, die evtl. schon im Filterschritt Antworten finden können, kann es in allen Schritten unserer Anfragebearbeitung deshalb nur darum gehen, nicht in Frage kommende Konstellationen zu verwerfen.

Abbildung 3 stellt die Architektur der Anfragebearbeitung dar, die aus folgenden Schritten besteht:

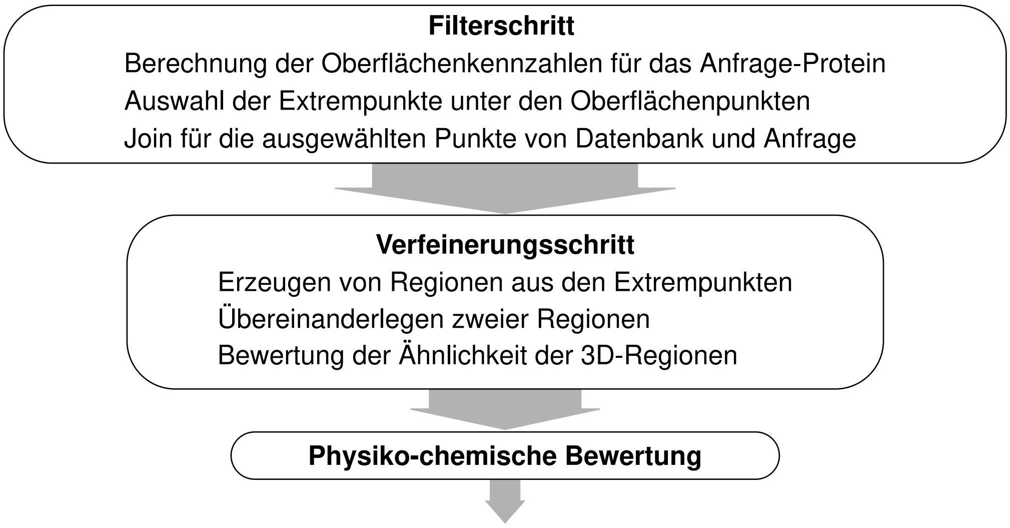

Abb. 3: Architektur der Anfragebearbeitung

- *Filterschritt mit geometrischen Kennzahlen*
  Für jeden Oberflächenpunkt der Proteine in der Datenbank werden verschiedene Kennzahlen berechnet, die die charakteristischen Eigenschaften der Geometrie in lageunabhängiger Weise darstellen. Von der großen Menge der Oberflächenpunkte werden diejenigen ausgewählt, bei denen die Oberflächenkennzahlen einen lokalen Extremwert annehmen. Bei Auswahl von ca. 50 von durchschnittlich 2000 Oberflächenpunkten eines Proteins erreichen wir eine Einschränkung des Konstellationsraums um den Faktor 40 * 40 = 1600, es verbleiben noch 6.25·106 potentielle Dockingkonstellationen. Die Kennzahlen der ausgewählten Punkte werden komplementiert und mit einem Verweis auf den Punkt abgespeichert. Zur eigentlichen Anfragebearbeitung werden die Kennzahlen für ein gegebenes Anfrageprotein berechnet. Der Filterschritt liefert mit Hilfe eines mehrdimensionalen Joins Paare aus je einem Punkt eines Proteins aus der Datenbank und einem Punkt des Anfrageproteins, deren Kennzahlen mindestens eine vorgegebene Ähnlichkeit besitzen.
- *Verfeinerungsschritt mit der 3D-Geometrie*
  Zum Vergleich der 3D-Geometrie werden räumlich ausgedehnte Objekte benötigt. Deshalb werden die vom Filterschritt gelieferten Extrempunkte beider Dockingpartner mit ihren Nachbarn und den verbindenden Kanten der Triangulierung zu Regionen zusammengefaßt. Die 3D-Geometrie der Regionen eines Kandidatenpaares wird nun miteinander verglichen, wofür die beste relative Position der zu vergleichenden Regionen im 3D zu bestimmen ist. Ergebnis des Verfeinerungsschritts sind Konstellationen von zwei Proteinen.
- *Bewertung der Konstellationen nach physikochemischen Eigenschaften*
  In den ersten beiden Schritten der Anfragebearbeitung werden nur die geometrischen Kriterien für Dockingkandidaten überprüft. Die fürs Docking ebenfalls relevanten physikalischen und chemischen Eigenschaften der Dockingregionen werden erst im letzten Schritt einbezogen, weil sie von der relativen Position zweier Proteine abhängen, die erst im Verfeinerungsschritt bestimmt wird.

4 Geometrische Oberflächenkennzahlen

Der Filterschritt der Anfragebearbeitung benutzt eine Beschreibung der geometrischen Oberflächencharakteristik mit Hilfe verschiedener Kennzahlen. Die Kennzahlen werden für jeden Punkt der Oberfläche berechnet und drücken geometrische Eigenschaften der Umgebung des Punktes aus. In der Literatur finden sich verschiedene Vorschläge für Kennzahlen. Für erste

Untersuchungen haben wir *Solid Angle* (SA) aus [Con 86a] sowie *Local Shape Index* (LSI) aus [Koe 90] gewählt, da sie sich in den Experimenten als gute Formbeschreibung erwiesen haben und einfach zu komplementieren sind. Die Auswahl einer für die Dockingsuche wirklich gut geeigneten Kombination von Kennzahlen ist eine wichtige, noch offene Frage.

Bei der zählenden Ermittlung des SA legt man eine Meßkugel K zugrunde, auf deren Oberfläche man eine bestimmte Anzahl n von Punkten gleichmäßig verteilt. Diese Kugel K legt man nun um P und zählt dann diejenigen Punkte auf K, die nicht im Inneren des Moleküls liegen (vgl. Abbildung 4). Teilt man das Ergebnis durch n und normiert es auf 4π, so erhält man ein Maß für den 3D-Öffnungswinkel ("Solid Angle") des Moleküls im Punkt P. Über den Radius der Meßkugel K hat man eine Skalierungsmöglichkeit für diese Kennzahl.

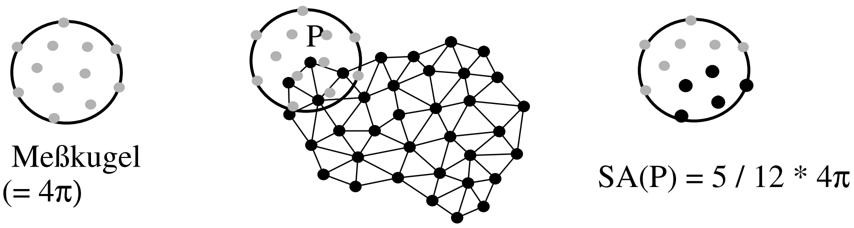

Abb. 4: Berechnung des SA für den Punkt P

Der LSI ist eine skalare Größe und beschreibt die Form der Oberfläche in der Umgebung eines Punktes P. Er kann als Maßzahl für die Konkavität bzw. Konvexität der Umgebung von P verstanden werden. Seine Werte liegen zwischen -1 und +1, die zugehörigen Formen sind in Abbildung 5 veranschaulicht.

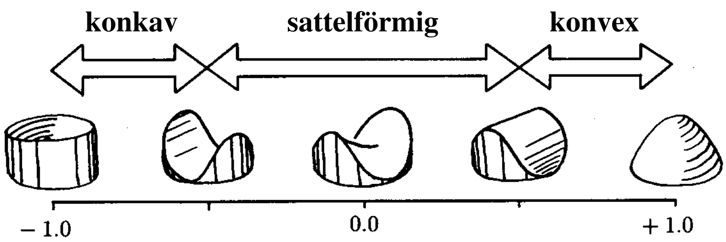

Abb. 5: Bedeutung des LSI (aus [Koe 90], S. 322)

Zur Berechnung des LSI am Punkt P wird ein elliptisches bzw. hyperbolisches Paraboloid in die Umgebung von P auf der triangulierten Oberfläche eingepaßt. Dann ermittelt man die beiden Hauptkrümmungen k_{min} und k_{max} als Eigenwerte der Hessematrix, die die zweiten Ableitungen des Paraboloids enthält, und faßt diese Krümmungswerte nach [Koe 90] zusammen.

Ein wichtiger Parameter für LSI ist der Selektionsabstand ("Radius") r, über den wir bestimmen, welche Nachbarn von P zum Einpassen berücksichtigt werden. Wir messen r entlang der Dreieckskanten und wählen dadurch neben den direkten Nachbarn eines Knotens auch weiter entfernt liegende Stützpunkte aus. Dadurch haben wir ein Maß, wie lokal oder regional die Form der Umgebung von P bestimmt werden soll. Für unsere Untersuchungen haben wir für r Werte zwischen 3 Å und 8 Å gewählt (Abbildung 6).

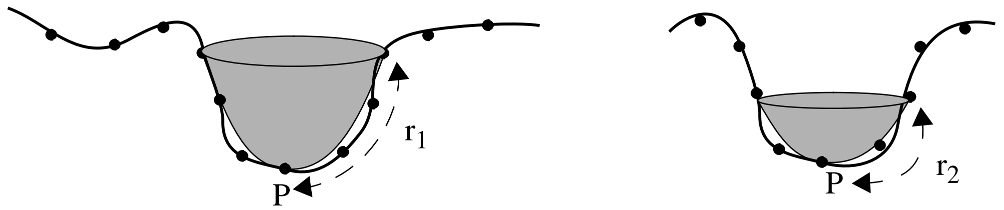

Abb. 6: Berechnung des LSI mit Selektionsabstand r_1 bzw. r_2

Für das Protein-Protein-Docking benötigen wir nicht Oberflächen, die ähnlich sind, sondern solche mit einer komplementären Form. Zur Vorbereitung der Anfragebearbeitung speichern wir die Kennzahlen der in der Datenbank gespeicherten Proteine komplementiert ab. Bei Annahme eines vernachlässigbaren Abstands der beiden Proteine an einer Dockingstelle können die Kennzahlen folgendermaßen komplementiert werden:

- Solid Angle: das Komplement erhält man als Differenz zum Maximalwert 4π.
- Local Shape Index: das Komplement ergibt sich durch Wechsel des Vorzeichens.

Mit SA und LSI haben wir nun zwei geometrische Kennzahlen, die die Form einer Oberflächenregion beschreiben und sich einfach komplementieren lassen. Sie werden im Filterschritt der Anfragebearbeitung eingesetzt, die im nächsten Kapitel beschrieben wird.

5 Die Schritte der Anfragebearbeitung

Der Filterschritt der Anfragebearbeitung, der auf den geometrischen Kennzahlen arbeitet, soll eine wesentliche Reduktion der Menge potentieller Konstellationen bewirken, ohne jedoch tatsächliche Dockingstellen auszuschließen. Zur effizienten Unterstützung des Filterschritts setzen wir eine mehrdimensionale Zugriffsstruktur, den R*-Baum [BKSS 90], ein. Beim Einfügen eines Proteins in die Datenbank wird folgende *Vorverarbeitung* durchgeführt:

- *Berechnung der Oberflächenkennzahlen.* Für jeden Oberflächenpunkt des in der Datenbank abzuspeichernden Proteins werden die Werte der Kennzahlen SA und LSI berechnet.
- *Auswahl der Extrempunkte der Oberfläche.* Alle Punkte der triangulierten Oberfläche, die bezüglich des SA in einer Umgebung von 6Å minimal oder maximal sind, werden ausgewählt. Dieses Vorgehen beruht auf der Annahme, daß die geometrisch signifikanten Regionen (konvexe oder konkave) beim Docking bedeutsam sind.
- *Update des R*-Baums für die Protein-Datenbank.* Die Kennzahlen der ausgewählten Punkte werden komplementiert. Die erhaltenen Punkte im k-dimensionalen Raum der Kennzahlen werden in einem R*-Baum abgespeichert und mit einem Verweis auf das Protein und die geometrische Lage innerhalb des Proteins versehen.

Die eigentliche *Anfragebearbeitung* läuft folgendermaßen ab:

- *Aufbau eines R*-Baums für das Anfrageprotein.* Für ein Anfrageprotein werden wie beim Einfügen in die Datenbank die Kennzahlen berechnet, die analog in einem zweiten R*-Baum verwaltet werden.
- *Join zwischen Datenbank und Anfrageprotein.* Ein mehrdimensionaler Join auf den beiden R*-Bäumen liefert alle Paare von je einem Punkt eines Proteins aus der Datenbank und einem Punkt des Anfrageproteins, deren Distanz im Kennzahlenraum kleiner als ein gegebenes ε ist. ε ist als minimaler Wert zu wählen, bei dem die bekannten Dockingstellen noch sicher gefunden werden.

Der Filterschritt wurde implementiert und mit Proteinen aus der PDB getestet. Dazu wurden Proteinkomplexe ausgewählt, die aus zwei oder mehreren Proteinen im gedockten Zustand bestehen. Sie wurden in ihre Teile zerlegt und in der Datenbank abgespeichert. Die Teilproteine wurden dann als Anfragen an die Datenbank gestellt. Diese Anfragen müssen in ihren Antwortmengen die aus der PDB bekannten Dockingpartner enthalten.

Abbildung 7 stellt die Anzahlen gefundener Punktpaare für die beiden Proteine 2ptc (Trypsin mit Trypsin Inhibitor) und 2pab (Präalbumin) bei verschiedenen Werten für ε dar.

Der Filterschritt mit allen Oberflächenpunkten hat erwartungsgemäß eine viel zu niedrige Selektivität, man erhält als Antwort etwa 20% aller Paare von Oberflächenpunkten. Bei Beschränkung auf die Extrempunkte der Oberfläche verbessert sich die Selektivität des Filterschritts um den Faktor 1000 bis 2000, was der Reduktion der Anzahl der Oberflächenpunkte um den Faktor 20 bis 40 etwa entspricht.

Protein	ε	Alle Punktepaare	Suche auf allen Punkten	Suche auf Extrempunkten	Punktepaare auf Dockingstelle
2ptc	1.5	$2.77 \cdot 10^6$	$0.53 \cdot 10^6$	236	1
	2.0		$0.73 \cdot 10^6$	341	2
	3.0		$1.17 \cdot 10^6$	545	2
2pab	1.5	$3.16 \cdot 10^6$	$0.46 \cdot 10^6$	617	0
	2.0		$0.66 \cdot 10^6$	760	1
	3.0		$1.13 \cdot 10^6$	1022	4

Abb. 7: Anzahlen gefundener Antwortpaare

Im Fall von 2pab wird die Dockingstelle bei ε = 1.5 nicht als Antwort geliefert, so daß man mit einem ε von mindestens 2.0 arbeiten muß. Eine Ursache für dieses relativ große ε liegt im bisherigen einfachen Verfahren zur Auswahl der Extrempunkte. Es treten nämlich Fälle auf, in denen die auf beiden Dockingpartnern gewählten Extrempunkte und damit auch ihre Kennzahlen relativ weit voneinander entfernt sind.

Der Filterschritt der Anfragebearbeitung liefert Paare von Punkten, die mit ihren Umgebungen für komplementäre Regionen stehen. Im Verfeinerungsschritte sind die Konstellationen explizit zu bestimmen sowie die geometrische Komplementarität der beiden Partner zu bewerten. Dazu sind folgende Aufgaben zu bearbeiten: Die Bildung von Regionen, die Normierung der Lage im Raum, die Bestimmung der relativen Positionen und die Berechnung einer Maßzahl für die Komplementarität.

Wir definieren die *Region* zu einem Punkt P als Menge von Oberflächenpunkten aus der Umgebung von P. Dazu wählen wir diejenigen Punkte aus, die von P aus über Kanten der Triangulierung innerhalb eines vorgegebenen Abstandes zu erreichen sind.

Die Region werden nun in eine normierte Lage gebracht, d.h. es wird eine Darstellung bestimmt, die translations- und rotationsinvariant ist. Für diese Normierung benötigen wir den Referenzpunkt R, den Normalenvektor N und eine Hauptachse H der Region. Alle drei Parameter werden durch die gewählten Approximationen zur Verfügung gestellt. Die Lage einer Region läßt sich durch sechs Parameter beschreiben: drei Koordinatenwerte für die Translation und drei Winkelwerte für die Rotation. Im ersten Schritt der Normierung werden die drei Freiheitsgrade der Translation bestimmt, im zweiten Schritt zwei Freiheitsgrade der Rotation. Zum Schluß wird eine Hauptachsentransformation durchgeführt, wodurch der dritte Freiheitsgrad der Rotation festgelegt wird (vgl. Abbildung 8).

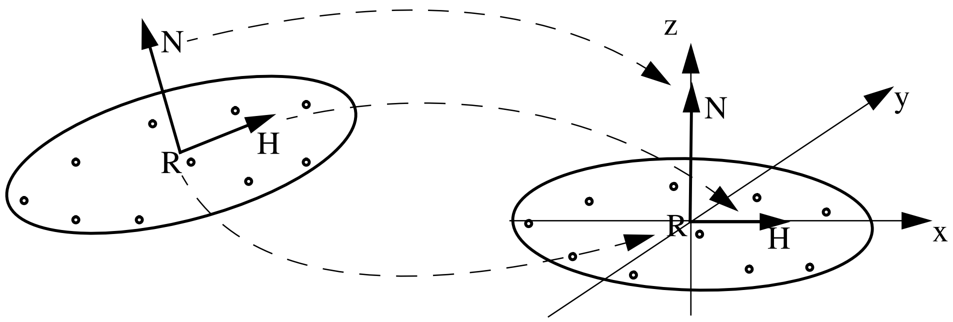

Abb. 8: Normierung der Lage einer Region

Mit Hilfe der normierten Lage wird die Konstellation genau bestimmt, d.h. die beiden Interaktionspartner werden zu einem Komplex zusammengesetzt. Dies geschieht dadurch, daß die beiden Regionen in ihrer normierten Lage übereinandergelegt werden.

6 Zusammenfassung und Ausblick

Das Protein-Docking ist eine Aufgabenstellung mit neuartigen Anforderungen an Datenbanksysteme. Die Objekte sind Körper im 3D mit komplexen Oberflächen. Das wesentliche Suchkriterium ist die Form der Objekte. Die absolute Lage und Orientierung der Objekte im Raum spielt—anders als bei herkömmlichen Anwendungen von Geo-Datenbanksystemen—für die Anfragebearbeitung im Dockingsystem keine Rolle.

Der prinzipiell unendlich große Suchraum für die Dockinganfrage eines Proteins an die Proteindatenbank wird durch die triangulierte Repräsentation der Oberflächen diskretisiert. Durch die Vorauswahl von Extrempunkten als potentielle Dockingstellen reduziert sich der gesamte Suchraum aller Punktpaare nochmals. Zur weiteren Einschränkung des Suchraums benutzen wir die bewährte Technik der mehrstufigen Anfragebearbeitung. Für den Filterschritt verwenden wir abstrakte Formbeschreibungen durch Kennzahlen. Die ausgewählten Punkte aller Proteine sowie diejenigen eines Anfrageproteins werden jeweils mit Hilfe eines R*-Baums verwaltet. Ein mehrdimensionaler Join liefert alle Paare von Punkten, deren Abstand im Kennzahlenraum eine vorgegebene Toleranz ε nicht überschreiten. Die Antworten aus dem Filterschritt sind Paare von Punkten, die im Verfeinerungsschritt zu Regionen erweitert werden. Diese Regionen werden in der Lage normiert und dann übereinandergelegt, um einfache Komplementaritätsprüfungen auf der 3D-Geometrie durchzuführen. Eine experimentelle Untersuchung des Filterschritts wurde beschrieben, für den Verfeinerungsschritt wird sie derzeit durchgeführt.

Wir haben verschiedene Komponenten des Dockingsystems vorgestellt, das wir zusammen mit unseren Partnern aus Braunschweig, Bielefeld und Göttingen im BMFT-Verbundprojekt BIOWEPRO entwickeln. Die Basis des Systems bildet ein kommerzielles objektorientiertes Datenbanksystem, das wir mit den Proteinen aus der PDB laden. Die Algorithmen zur Berechnung der Oberflächenrepräsentation, der Oberflächenkennzahlen und der potentiellen Dockingstellen sowie die Module der mehrstufigen Anfragebearbeitung werden zusammen mit den Bausteinen unserer Projektpartner zu einem Gesamtsystem integriert.

Danksagung

Wir danken unserem Kollegen Ralf Schneider für intensive und fruchtbare Diskussionen. Unseren Partnern im Projekt BIOWEPRO verdanken wir insbesondere die Einführung in die biologische Problemstellung.

Literaturhinweise

[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, Evanston, ILL, in: Lecture Notes in Computer Science, Vol. 730, Springer, 1993, pp. 69-84.

[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', 05.-07.09.94, Jena, 1994.

[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.

[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, 1993, in: Informatik aktuell, Springer, 1993, pp. 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.

[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.

[BKSS 94] Brinkhoff T., Kriegel H.-P., Schneider R., Seeger B.: *'Efficient Multi-Step Processing of Spatial Joins',* Proc. ACM SIGMOD Int. Conf. on Management of Data, Minneapolis, MN, 1994, pp. 197-208.

[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 86a] Connolly M. L.: *'Measurement of protein surface shape by solid angles'*, Journal of Molecular Graphics, Vol. 4, No. 1, 1986, pp. 3-6.

[Con 86b] 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 Symposium on Combinatorial Pattern Matching (CPM '93), Padova, Italy, in: Lecture Notes in Computer Science, Vol. 684, Springer, 1993, pp. 20-34.

[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.

[HT 94] Helmer-Citterich M., Tramontano A.: *'PUZZLE: A New Method for Automated Protein Docking Based on Surface Shape Complementarity',* Journal of Molecular Biology, Vol. 235, 1994, pp. 1021-1031.

[Jag 91] Jagadish H. V.: *'A Retrieval Technique for Similar Shapes',* Proc. ACM SIGMOD Int. Conf. on Management of Data, Denver, CO, 1991, pp. 208-217.

[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.

[Koe 90] Koenderink J. J.: *'Solid Shape',* MIT Press, Cambridge, MA, 1990.

[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.

[MG 93] Mehrotra R., Gary J. E.: *'Feature-Based Retrieval of Similar Shapes',* Proc. 9th Int. Conf. on Data Engineering, Vienna, Austria, 1993, pp. 108-115.

[PD 93] Pellegrini M., Doniach S.: *'Computer Simulation of Antibody Binding Selectivity',* Proteins: Structure, Function, and Genetics, Vol. 15, 1993, pp. 436-444.

[PDB 94] Protein Data Bank: *'Quarterly Newsletter No. 70 (October 1994)',* Brookhaven National Laboratory, Upton, NY, 1994.

[Ric 77] Richards F. M.: *'Areas, Volumes, Packing, and Protein Structure',* Annual Reviews in Biophysics and Bioengineering, Vol. 6, 1977, pp. 151-176.

[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.

[Ull 88] Ullman J. D.: *'Principles of Database and Knowledge-Base Systems (Volume 1)',* Computer Science Press, Rockville, ML, 1988.