Veröffentlicht im Tagungsband: “Bioinformatik – Computereinsatz in den Biowissenschaften”
9. Workshop “Computereinsatz in den Biowissenschaften”
2. GI-Fachtagung “Informatik in den Biowissenschaften”, Jena, 05.- 07.09.94

Datenbankunterstützung für das Protein-Protein-Docking: ein effizienter und robuster Feature-Index[1]

Kai Aldinger, Martin Ester, Gabriele Förstner, Hans-Peter Kriegel, Thomas Seidl

Institut für Informatik, Universität München

Leopoldstr. 11B, D-80802 München

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

Zusammenfassung

In der vorliegenden Arbeit schlagen wir eine Architektur für ein Dockingsystem vor, das zu einem gegebenen Anfrageprotein alle möglichen Dockingpartner und deren zugehörige Konstellationen in einer Proteindatenbank sucht. Die einzelnen Filter- und Bewertungsschritte dieses Dockingsystems sollen durch den Einsatz räumlicher Zugriffsstrukturen unterstützt werden. Hier behandeln wir insbesondere die Verwendung eines effizienten und robusten Feature-Index. Dieser dient dazu, aus dem Raum aller möglichen Konstellationen von Partnerproteinen aus der Datenbank gute Kandidatenvorschläge zu ermitteln. Dazu werden im Vorfeld die Proteinoberflächen regionalisiert, in einem k-dimensionalen Raum von Kennzahlen beschrieben und in einer k-dimensionalen räumlichen Indexstruktur verwaltet. Für die Docking-Suche wird das Anfrageprotein analog regionalisiert, und zusätzlich werden seine Kennzahlen komplementiert, so daß zum Komplement ähnliche Regionen als mögliche Docking-Stellen im Index gefunden werden können. In den nachfolgenden Filterschritten des Dockingsystems wird unter anderem das räumliche Passen genauer überprüft.

Schlüsselwörter: Docking-Suche auf Protein-Datenbanken, räumliche Zugriffsstrukturen, Anfragebearbeitung in räumlichen Datenbanksystemen.

1. Einleitung

Die Funktionsweise globulärer Proteine wie Enzyme, Inhibitoren, Repressoren u.a. besteht grundsätzlich in der Wechselwirkung mit kleinen Liganden, mit anderen Proteinen sowie mit sehr großen Biomolekülen (DNA). Diese Dockingvorgänge werden wesentlich durch verschiedene geometrische und nichtgeometrische Oberflächeneigenschaften wie etwa Krümmung, Ladung und Hydrophobie gesteuert.

Im Rahmen des BMFT-Verbundprojekts BIOWEPRO (Biomolekulare Wechselwirkungen von Proteinen) mit der GBF Braunschweig, der Universität Bielefeld und dem MPI Göttingen soll ein Protein-Protein-Docking-Datenbanksystem entwickelt werden, das Biologen bei der Suche von geeigneten Dokking-Kandidaten unterstützt. Im Gegensatz zu den meisten bisher veröffentlichten Ansätzen soll dabei nicht nur ein Protein-Paar auf gegenseitige Docking-Möglichkeit (*1:1-Docking*) untersucht werden, sondern ein gegebenes Protein mit allen Proteinen in der Datenbank (*1:n-Docking*). Zu einem als Docking-Anfrage auf der Proteindatenbank [Ber 77] gegebenen Molekül soll das System auf die folgenden beiden Fragen eine Antwort finden:

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.

- Welche Proteine der Datenbank kommen als Dockingkandidaten in Frage (Kandidatenauswahl) und welche Proteine scheiden sicher aus (Negativauswahl)?
- An welcher Position bzw. in welchem Bereich kann ein vorgeschlagenes Protein geometrisch und energetisch günstig mit dem Anfrage-Molekül wechselwirken?

Von praktischer Bedeutung sind diese Fragen u.a. in der Pharmazie, beispielweise zur Untersuchung von Abwehrreaktion des Immunsystems auf Fremdkörper, zur Überprüfung von Medikamenten auf Nebenwirkungen oder bei der Suche nach einem Impfstoff zu einem aktuell grassierenden Grippevirus. Gelingt es durch den Einsatz eines Docking-Systems die pharmazeutische Forschung bei der Suche nach Antworten auf diese Fragen zu unterstützen, läßt sich durch ein zielgerichtetes Vorgehen bei der Entwicklung neuer Medikamente die Anzahl aufwendiger Laborversuche reduzieren und somit auch die Zahl der nötigen Tierversuche einschränken [PBEÖ 90].

Es gibt zwei grundlegende Ansätze zum Finden von Docking-Kandidaten [Len 93]. Beim ersten Ansatz wird eines der beiden Moleküle so verschoben und rotiert, daß es zu dem anderen paßt [WS 92] [Kat 92]. Wegen der großen Anzahl verschiedener Lagen, die die beiden Moleküle zueinander einnehmen können und die somit probiert werden müssen, sind diese Verfahren schon für einen einzigen Docking-Kandidaten sehr aufwendig und kommen wohl für die Suche nach allen Docking-Kandidaten in einer Datenbank nicht in Frage; dieser Ansatz eignet sich häuptsächlich für das 1:1-Docking. Beim zweiten Ansatz dagegen wird ein abstraktes Modell der Protein-Oberflächen berechnet und in der Datenbank gespeichert, das die Definition eines Maßes für die Ähnlichkeit zweier Oberflächenstücke erlaubt [Con 86a][SBK 92]. Mit Hilfe solcher Verfahren läßt sich die Zahl der fürs Docking in Frage kommenden Proteine bzw. Regionen effizient einschränken; sie sind somit für das 1:n-Docking einsetzbar. Aufgrund des Informationsverlusts bei der Modellbildung können jedoch auch Kandidaten geliefert werden, die nicht docken. Es ist deshalb in einem späteren Schritt nötig, die Gesamtstruktur der gefundenen Kandidaten mit der des Anfrage-Proteins geometrisch exakt zu vergleichen.

In diesem Papier verfolgen wir den zweiten Ansatz. Im Folgenden wird zunächst eine mögliche Einbettung des Feature-Index in ein vollständiges Docking-System vorgestellt. Einen kurzen Überblick über die von uns verwendeten Kennzahlen zur Beschreibung der Protein-Oberflächen gibt Abschnitt 3. In Abschnitt 4 wird dann ein Verfahren zur Regionalisierung von Protein-Oberflächen angegeben und in Abschnitt 5 die Anfragebearbeitung zur Suche nach komplementären Regionen vorgestellt. Die Implementierung des Feature-Index, der es erlaubt, die Suche nach Dockingstellen auf diejenigen Oberflächenregionen von Molekülen zu beschränken, die zum Anfrage-Protein (bzw. einer seiner Oberflächenregionen) komplementäre Oberflächenkennzahlen besitzen, erfolgt schließlich in Abschnitt 6. Dieser Feature-Index ist robust, d.h. er findet nicht nur exakt komplementäre, sondern auch zum Komplement ähnliche Features.

2. Gesamtarchitektur des Docking-Systems

In diesem Kapitel wird ein Überblick über eine mögliche Architektur für ein Docking-System gegeben, wobei wir hauptsächlich auf den aus Datenbanksicht relevanten Einsatz von Zugriffsstrukturen innerhalb des Systems eingehen. Die dem Docking-System zugrunde liegende Idee ist es, zu einem gegebenen Protein die Menge der dem System bekannten Proteine zu finden, die mit einem gegebenen Protein docken, also biochemische Wechselwirkungen eingehen. Zusätzlich soll das System zu jedem von ihm gefundenen Protein die relative Lage liefern, in der es mit dem gegebenen Protein dockt. Die relative räumliche Lage zweier Proteine zueinander wird im folgenden als Protein-Protein-Konstellation oder kurz als *Konstellation* bezeichnet. Da sich diese Menge in der Realität jedoch nur eingrenzen, aber nicht hundertprozentig bestimmen läßt, müssen die vom Docking-System gelieferten Resultate letztendlich vom Molekularbiologen im Experiment überprüft werden.

Ein wesentliches Architekturmerkmal des hier vorgestellten Docking-Systems ist die Bearbeitung einer Anfrage in mehreren Schritten. In einem ersten Schritt wird über Regionen auf den Proteinoberflächen eine möglichst kleine Menge von in Frage kommenden Protein-Protein-Konstellationen gesucht, die dann in den folgenden Schritten nach unterschiedlichen chemischen und geometrischen Kriterien bewertet und reduziert wird. Diese Mehrschrittbearbeitung hat sich in räumlichen Datenbanksystemen für die Bearbeitung von Anfragen [KSB 93] bzw. von Operationen wie dem Spatial Join [BKSS 94] als sehr effizient erwiesen. Besonders wegen der anfänglich sehr großen Menge potentieller Konstellationen wird in den ersten Schritten auf ihre genaue Untersuchung zu Gunsten geringerer Kosten pro Konstellation verzichtet. Im Vordergrund steht zunächst vielmehr, Konstellationen, die mit Sicherheit nicht in Frage kommen, möglichst frühzeitig auszusortieren und für diese somit weitere, mit Kosten verbundene, unnötige Bewertungen zu vermeiden. Diese Vorgehensweise basiert auf der Annahme, daß sich viele der Konstellationen schon mit einfachen Mitteln von der Menge der potentiellen Konstellationen ausschließen lassen. Für andere hingegen ist eine genauere Bewertung nötig, bevor entschieden werden kann, ob diese Konstellation sinnvoll ist.

Die Oberflächen der Proteine werden zunächst regionalisiert und im ersten Schritt mit Hilfe der dadurch entstandenen Regionen Protein-Protein-Konstellationen gesucht. Durch eine normalisierte geometrische Darstellung der Regionen oder durch eine Charakterisierung der Regionen durch bestimmte Kennzahlen wird eine translations- und rotationsunabhängige Darstellung der Proteinoberflächen innerhalb des Systems erreicht. Das gegebene Anfrage-Protein wird ebenfalls regionalisiert und für die dabei entstandenen Regionen nach passenden Gegenstücken in der Datenbank gesucht. In den folgenden Bewertungsschritten werden die gefundenen Konstellationen dann genauer bewertet und nicht in Frage kommende Konstellationen verworfen. Wegen der immer kleiner werdenden Menge von potentiellen Dockingstellen dürfen die späteren Schritte durchaus auch mit höheren Kosten pro Dockingkandidaten verbunden sein.

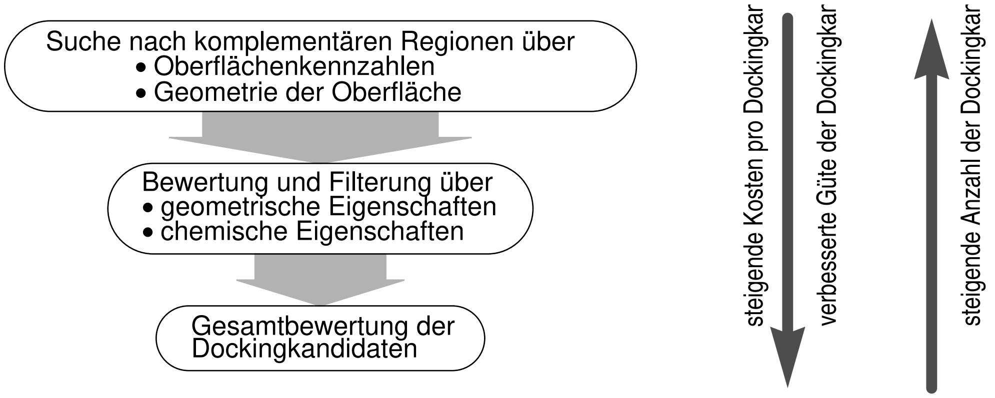

Abb. 1: Überblick über die Architektur des Docking-Systems

Abbildung 1 gibt einen groben Überblick über die datenbankrelevanten Komponenten des von uns vorgeschlagenen Docking-Systems. Es läßt sich dabei eine grobe Einteilung in drei Schritte vornehmen:

- *Finden möglicher Protein-Protein-Konstellationen*
  Zu den durch die Regionalisierung der Oberfläche des gegebenen Proteins entstandenen Regionen wird nach komplementären Oberflächenregionen auf der Menge der im Dockingsystem bekannten Proteine gesucht. Die prinzipielle Schwierigkeit bei dieser Vorgehensweise liegt darin, daß eine rotations- und translationsunabhängige Repräsentation für die Regionen auf den Proteinoberflächen gefunden werden muß. Zwei Ansätze werden von uns in diesem Zusammenhang verfolgt: Die Repräsentation der Regionen auf den Proteinoberflächen durch Kennzahlen und durch eine Approximation der exakten Geometrie der Oberflächenregionen. Der Ansatz über die Kennzahlen wird in den

folgenden Kapiteln noch ausführlicher erörtert. Beim zweiten Ansatz, den wir in diesem Artikel nicht weiter behandeln, wird nach Oberflächenregionen gesucht, die dem Komplement der Anfrageregion in der dreidimensionalen Geometrie möglichst ähnlich sind. Dazu ist es notwendig, die Geometrie durch Normalisierung in eine translations- und rotationsinvariante Darstellung zu bringen.

- *Bewertung und Filterung der Protein-Protein-Konstellationen*
  Für die Suche in den nachfolgenden Schritten werden die gefundenen Dockingstellen nach verschiedenen Kriterien genauer bewertet und nicht in Frage kommende Dockingkandidaten möglichst früh herausgefiltert. Die Bewertung erfolgt dabei über die geometrischen Eigenschaften der Proteinoberflächen sowie über die Evaluierung der biochemischen Wechselwirkungen.
  Die Filter, die schlechte Dockingstellen verwerfen, müssen pessimistisch ausgelegt werden, d.h. sie dürfen keine Konstellationen entfernen, die in der Realität Dockingkandidaten sein können. Auf der anderen Seite muß die Selektivität jedes Schrittes hoch genug sein, um die Menge der in Frage kommenden Dockingkandidaten erheblich einzugrenzen. In dieser Hinsicht besteht zwischen diesen beiden Zielen ein Konflikt.
  Auch muß von der lokalen Betrachtung der Proteinoberflächen (Regionen) auf die globale Betrachtung der Proteine selbst übergegangen werden; Konstellationen, bei denen sich die beiden beteiligten Proteine an irgendeiner Stelle räumlich stark überschneiden, kommen beispielsweise nicht mehr als Dockingkandidaten in Betracht, auch wenn sie lokal gut dockende Regionen besitzen.
- *Gesamtbewertung der Protein-Protein-Konstellationen*
  Die Dockingkandidaten mit ihren durch die Filterschritte gewonnenen Ergebnissen werden durch das Docking-System verwaltet. Aus den Teilbewertungen der einzelnen Konstellationen soll eine Gesamtbewertung (Ranking) aller Dockingkandidaten erstellt werden. Eine einfache Möglichkeit hierzu ist es, die gewichtete Summe der Teilbewertungen zu bilden und die Kandidaten danach zu sortieren. Alternativ ist z.B. der Einsatz heuristischer Regeln denkbar.

In den folgenden Kapiteln wird nun die Repräsentation der Oberfläche mittels Kennzahlen und der darauf basierende Feature-Index genauer behandelt.

3. Oberflächenkennzahlen für Proteine

Die Oberfläche der Moleküle definiert man in der Regel über die Approximation der Atome durch Kugeln mit den zugehörigen van der Waals-Radien. Neben dem einfachen van der Waals-Modell gibt es auch geglättete Formen der molekularen Oberfläche. [Ric 77] beschreibt eine Form, die aus konvexen, sattelförmigen und konkaven Stücken besteht und über die Zugänglichkeit durch eine Probenkugel definiert ist. Die Exaktheit dieser Repräsentation benötigt man in späteren Filterschritten, um räumliches Passen und Überschneidungen feststellen zu können. Für die frühen Phasen der Docking-Suche ist sie jedoch wegen ihrer Komplexität aufgrund der Unterschiedlichkeit der Komponenten wenig geeignet. Wir verwenden deshalb für den Feature-Index eine Triangulation von Oberflächenpunkten des Moleküls. Dazu wird eine bestimmte Anzahl von Punkten jeweils gleichmäßig auf den Atomschalen verteilt (hier etwa 6, 8, 10, 12 oder 14 Punkte als Ecken von regelmäßigen Polyedern). Mit Hilfe des von uns entwickelten Programms CAPS (CAlculation of Protein Surfaces) [Sch 94] wird eine Triangulation derjenigen Punkte berechnet, die durch die Probenkugel zugänglich sind; die nicht-zugänglichen Punkte werden dabei eliminiert. Man erhält somit eine geschlossene, geglättete und handliche Oberflächenrepräsentation.

In der Literatur finden sich verschiedene Vorschläge zur Charakterisierung von Proteinoberflächen mit Hilfe von *Kennzahlen*, z.B. “Density of Surface Neighborhood (DSN)” [CM 90], “Solid Angle (SA)” [Con 86b], “Shape Distribution” [Con 92], “Shape Index” [DO 93], “Global Curvature” [ZHSB 92], “Topologische Einbettung (Embeddedness, EMB)” und “Surface Topology Index (STI)” [Hei 93]. Dabei wird jeweils zu einem Oberflächenpunkt P eine charakteristische Größe über die Beschaffenheit seiner

Umgebung ermittelt. Ein solcher Bezugspunkt kann z.B. ein Voxel (bei Digitalisierung der Oberfläche) oder ein Eckpunkt eines Dreiecks (bei Triangulierung der Oberfläche) sein. Die Berechnung erfolgt entweder zählend (DSN, Solid Angle, Shape Distribution), durch eine einfache Summenformel (EMB) oder differentialgeometrisch (Shape Index, Global Curvature, STI).

Die Werte für EMB erhält man beispielsweise für jedes P durch eine einfache Iteration über alle anderen Atome a_i, wobei jeweils der Abstand $d_i = d(P, a_i)$ reziprok in eine Summenformel eingeht.

Die differentiellen Ansätze basieren auf der Berechnung der beiden Hauptkrümmungen der Oberfläche am Punkt P, die man als Eigenwerte der lokalen Hesse-Matrix erhält. Die Hesse-Matrix setzt sich aus den zweiten Ableitungen der Oberfläche zusammen, wozu man eine zweimal differenzierbare Darstellung der Oberfläche benötigt. In [DO 93] werden dazu Gauß-Funktionen verwendet, in [Hei 93] hyperbolische und elliptische Paraboloide in die Umgebung von P eingepaßt. Beide Verfahren gehen von einer triangulierten Repräsentation der Oberfläche aus.

Neben den geometrischen Kennzahlen kann man der Oberfläche auch nicht-geometrische Kennzahlen zuordnen. Dazu eignen sich etwa die Partialladung, der Potentialwert oder der Grad der Hydrophobie im Punkt P. Diese Eigenschaften werden jeweils durch die nächstliegenden Atome induziert, wofür der Bezug zwischen Oberflächenpunkten und den benachbarten Atomen erforderlich ist.

Für erste Untersuchungen haben wir Solid Angle (zählende Ermittlung) sowie eine modifizierte Form des STI (analytische Berechnung) ausgewählt, da sie beide einfach zu implementieren und zudem anschaulich zu erklären sind. Die Auswahl einer für die Dockingsuche wirklich gut geeigneten Kombination von Kennzahlen ist eine wichtige und hier noch offene Frage.

Bei der zählenden Ermittlung des "Solid Angle" 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. Teilt man das Ergebnis durch n, 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.

Die Berechnung des STI am Punkt P erfolgt in zwei Schritten: Im ersten Schritt wird ein elliptisches bzw. hyperbolisches Paraboloid in die triangulierte Oberfläche eingepaßt. Dies geschieht durch die Methode der kleinsten Fehlerquadrate, wobei die Nachbarknoten von P als Stützpunkte verwendet werden. Als Möglichkeit zur Skalierung des STI kann man neben den direkten Nachbarn eines Knotens auch weiter entfernt liegende Knoten als Stützpunkte wählen. Die Auswahl läßt sich über einen vorgegebenen Selektionsabstand steuern, der entlang der Dreieckskanten gemessen wird.

Im zweiten Schritt ermittelt man die beiden Hauptkrümmungen des Paraboloids und faßt sie zur skalaren Größe STI zusammen, die als Maßzahl für die Konkavität bzw. Konvexität verstanden werden kann. Die STI-Werte liegen zwischen 0 und 4, wobei man zur Veranschaulichung den ganzen Zahlen folgende Begriffe zuordnen kann [Hei 93]: 0 – Beutel, 1 – Spalt, 2 – Sattel, 3 – Grat, 4 – Pfropf.

Die angegebenen Oberflächenkennzahlen beschreiben zwar jeweils Umgebungen, sind aber dennoch punktbezogen definiert. Punkte, die bezüglich der verwendeten Oberflächenkennzahlen ähnlich sind, lassen sich zu Regionen zusammenfassen.

Für die Verwendung im Docking-System kommt eine wichtige Anforderung an die Kennzahlen hinzu, nämlich ihre Komplementierbarkeit. Bei der Docking-Suche sind alle Regionen von Proteinen in der Datenbank zu ermitteln, die zu einer Region des Anfrageproteins komplementär, d.h. zum Komplement einer Anfrageregion ähnlich sind.

Die Anforderung nach Komplementierbarkeit ist bei STI gut erfüllt: Ein Pfropf paßt auf einen Beutel, ein Grat in einen Spalt, zwei Sattelflächen können aufeinanderpassen. Vorausgesetzt, die Größe stimmt überein, läßt sich die Komplementierung von STI also anschaulich durch Ergänzung auf 4.0 berechnen.

Bei “Solid Angle” ermittelt man das Komplement ähnlich einfach durch Ergänzung auf die volle Anzahl an Punkten auf der Meßkugel. Für das Maß EMB stößt man hier auf Schwierigkeiten, ein Komplement-Begriff läßt sich schwer vorstellen. Das liegt vor allem darin begründet, daß EMB weniger form- als vielmehr lageabhängig bezüglich des Molekülschwerpunktes ist.

Wie erwähnt bleibt zu klären, welche Oberflächenkennzahlen sich für die Docking-Suche eignen. Wichtige Kriterien dafür sind, daß die auszuwählenden Kennzahlen möglichst voneinander unabhängig, einfach und eindeutig komplementierbar sind und durch eine gute Charakterisierung unterschiedlicher Formen eine hohe Selektivität sicherstellen.

4. Regionalisierung und Abspeicherung von Regionen

Dockingstellen sind nicht punktförmig, sondern Teilflächen der Oberfläche eines Proteins. Wir wollen die in Kapitel 3 vorgestellten Oberflächenkennzahlen nutzen, um charakteristische und für das Docking relevante Flächen zu bestimmen. Als *Region* bezeichnen wir eine Menge von benachbarten Punkten auf der Proteinoberfläche.

Wir gehen im folgenden davon aus, daß für alle Oberflächenpunkte eines Proteins k verschiedene Kennzahlen (in unserem Beispiel STI und SA, also k = 2) berechnet sind. Den durch die Kennzahlen definierten k-dimensionalen Raum bezeichnen wir als *Featureraum*. Ein Tupel (Kennzahl$_1$, . . ., Kennzahl$_k$) bezeichnen wir als *Feature*. Punkte der Proteinoberfläche, die ähnliche Features besitzen, sollen zu Regionen zusammengefaßt werden. Abbildung 2 zeigt ein Beispiel für die Regionalisierung von Proteinoberflächen aufgrund der Oberflächenkennzahlen STI und Solid Angle.

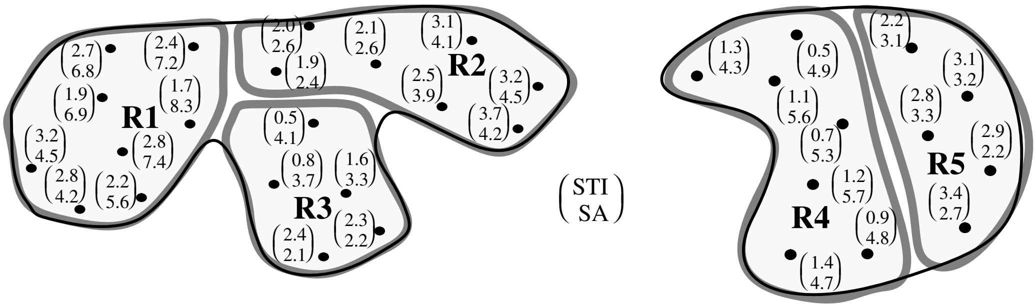

Abb. 2: Beispiel für die Regionalisierung von Proteinoberflächen

Ein Verfahren für eine solche Regionalisierung ist das Regionenwachstum [Nie 90]. Man beginnt mit punktförmigen Regionen und erweitert diese solange um benachbarte Punkte, wie die Ähnlichkeit dieser Punkte zur bisherigen Region hinreichend groß ist. Diese Methode verwendet [Hei 93] zur Regionalisierung von Proteinoberflächen. Im Unterschied zu unserem Ansatz benutzt er jedoch nur eine Kennzahl. Wir verfolgen einen anderen Ansatz zur Regionalisierung, der durch die beabsichtigte Abspeicherung mit Hilfe einer mehrdimensionalen Zugriffstruktur zur effizienten Anfragebearbeitung motiviert ist. Eine solche Zugriffstruktur faßt die Objekte im Datenraum so in mehrdimensionale, *minimal umgebende achsenparallele Rechtecke (MUR)* zusammen, daß diese Rechtecke ähnliche Objekte enthalten und möglichst disjunkt sind. Wir betrachten nun als Datenraum den um die drei geometrischen Dimensionen erweiterten Featureraum (Dimension in unserem Fall gleich 5). Die MUR’s in diesem Datenraum definieren Regionen auf der Proteinoberfläche, denn alle Punkte in einem solchen Rechteck besitzen ähnliche Werte für die k Kennzahlen und sind räumlich benachbart.

Zum Suchen komplementärer Regionen berücksichtigen wir nur die translations- und rotationsinvarianten Kennzahlen der Oberflächenpunkte, nicht aber ihre Koordinaten im 3D-Raum. Die Suche erfolgt also im Featureraum. Im Feature-Index speichern wir durch MUR’s zusammengefaßte Features ab und geben jedem solchen MUR einen Verweis auf die zugehörige Region einer Proteinoberfläche mit. Abbildung 3 stellt diese Abspeicherung der Features und der Verweise auf die Regionen mit Hilfe einer mehrdimensionalen Zugriffsstruktur dar.

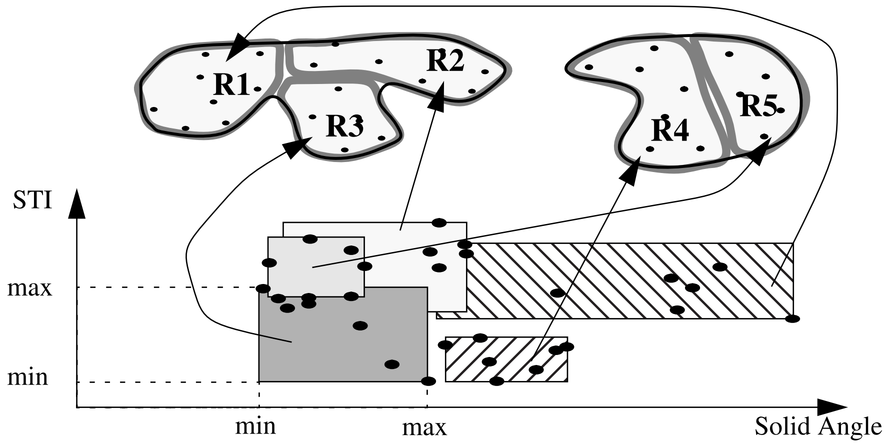

Abb. 3: Abspeicherung der Features

5. Anfragebearbeitung zum Suchen komplementärer Regionen

Proteine, die Anfrage-Proteine für die Datenbank sind, werden genau wie die in der Datenbank abgespeicherten Proteine behandelt: für jeden Punkt werden die Oberflächen-Kennzahlen berechnet, und nach deren Werten wird die Oberfläche des Anfrage-Proteins regionalisiert. Danach werden die Oberflächen-Kennzahlen des Anfrage-Proteins komplementiert. Eine Menge komplementierter Features beschreibt eine Region, die zur ursprünglichen Region komplementär ist. Die Komplementierung von Feature-Mengen bzw. von Regionen veranschaulicht Abbildung 4. Über den Feature-Index werden schließlich Regionen gesucht, die einer komplementierten Region des Anfrage-Proteins ähnlich sind.

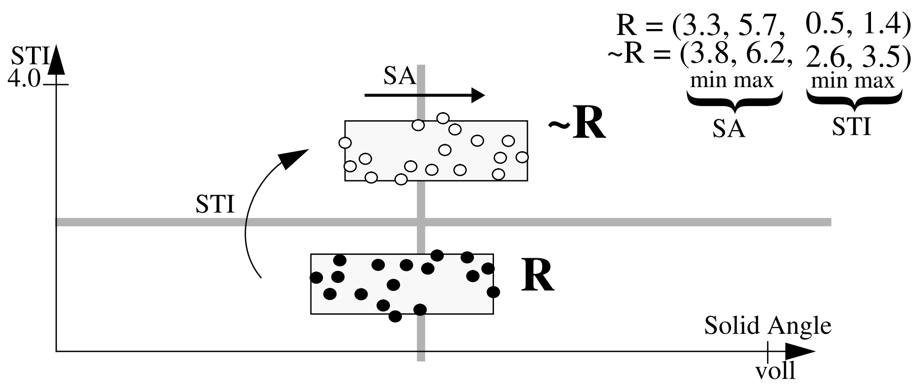

Abb. 4: Komplementierung von Feature-Mengen

Wir schlagen im folgenden ein Verfahren vor, das zu einer gegebenen Region alle ähnlichen Regionen über den Feature-Index liefert. Diese Anfragebeantwortung geschieht in zwei Schritten:

Filterschritt

Im ersten Schritt werden nur die MUR's der Features der Datenbank-Regionen und der Anfrage-Region miteinander verglichen. Wenn sie sich schneiden, dann ist ein Kandidat gefunden, sonst ist die zugeordnete Region sicher keine Antwort. Wenn nämlich die MUR's zweier Feature-Mengen disjunkt sind, dann sind die Feature-Mengen selbst auch disjunkt und damit die beiden zugehörigen Regionen sicher nicht ähnlich. Umgekehrt sind die vom Filterschritt gefundenen Kandidaten jedoch noch weiter zu untersuchen. Es kann nämlich der Fall auftreten, daß die MUR's von zwei Feature-Mengen (von einer gespeicherten Region und der Anfrage-Region) sich schneiden, ohne daß in diesem Überlappungsgebiet ein einziges Feature der beiden Regionen liegt, siehe dazu das Beispiel in Abbildung 5.

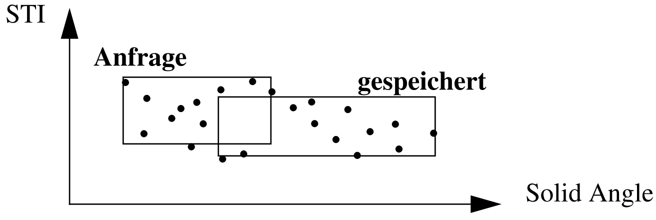

Abb. 5: Beispiel zur Notwendigkeit des Verfeinerungsschritts

Verfeinerungsschritt

Im zweiten Schritt werden für die gefundenen Kandidaten die Features selbst mit denen der Anfrage-Region verglichen, wobei wir uns auf die Features im Überlappungsgebiet beschränken können. Dieser Vergleich liefert den Grad der *Ähnlichkeit* der Features der Regionen zu denen der Anfrage-Region. Welche Ähnlichkeitsmaße hierzu besonders geeignet sind, bleibt noch zu untersuchen. Bei Wahl geeigneter Kennzahlen sind die Regionen ähnlicher Feature-Mengen selbst auch ähnlich. Folgende Ähnlichkeitsmaße für Feature-Mengen sollen beispielsweise untersucht werden:

- die Differenz der Kardinalitäten der beiden Feature-Mengen: bei Annahme der Gleichverteilung der Oberflächenpunkte im 3D-Raum liefert die Anzahl von Punkten einer Region und somit auch die Anzahl von Elementen der zugehörigen Feature-Menge ein Maß für die Größe dieser Region.
- der Abstand der Schwerpunkte der beiden Feature-Mengen: mit Hilfe der Schwerpunktbildung können Ausreißer ignoriert werden.

Die Regionen aller Feature-Mengen, deren Ähnlichkeit zur Feature-Menge der Anfrageregion größer als ein bestimmter Schwellwert ist, werden als Antwort auf die Anfrage ausgegeben. Durch die Bewertung der Ähnlichkeit erhält man eine Ordnung der Antworten, die nachfolgende Schritte im Docking-System nutzen können, indem sie z.B. diese Bewertung als Teil einer Gesamtbewertung der Docking-Kandidaten nehmen oder die Antworten des Feature-Index in der erhaltenen Ordnung (d.h. die besten zuerst) weiterverarbeiten.

Auch die Antworten des Verfeinerungsschritts sind jedoch noch nicht mit Sicherheit Docking-Kandidaten. Der Grund dafür liegt darin, daß der Feature-Index die Lage eines Oberflächenpunkts im 3D-Raum vernachlässigt, um eine translations- und rotationsunabhängige Darstellung zu erreichen. Es kann nun im schlechtesten Fall sein, daß zwei Regionen zwei gleiche Features besitzen, die zugehörigen Oberflächenpunkte jedoch auf der einen Region einen kleinen und auf der anderen Region einen großen Abstand voneinander besitzen. In diesem Fall könnte die gefundene Antwort-Region nicht mit der Anfrage-Region docken. Die vom Feature-Index gelieferten Regionen müssen also in jedem Fall noch mit der Anfrage-Region auf geometrische Ähnlichkeit geprüft werden. Dies wird die Aufgabe für einen weiteren Schritt des in Kapitel 2 skizzierten Docking-Systems sein.

Abbildung 6 stellt die Bearbeitung der Anfrage mit der zu R komplementären Feature-Menge ~R auf den Regionen bzw. deren Feature-Mengen aus Abbildung 3 dar. R3, R4 und R5 sind sicher keine Antworten, da ihre MUR's das MUR der Feature-Menge ~R nicht schneiden (Filterschritt). R1 hat 14, R2 hat 6 Feature-Punkte im Überlappungsgebiet mit ~R. R1 ist also ähnlicher zu ~R als R2 (Verfeinerungsschritt),

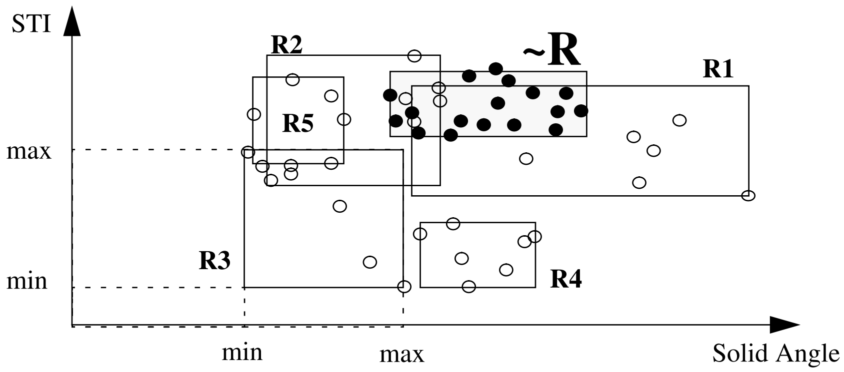

Abb. 6: Anfragebearbeitung für Anfrage ~R auf den Feature-Mengen aus Abbildung 3

6. Implementierung mit Hilfe des R*-Baums

Um auf große Mengen von Daten effizient zuzugreifen bzw. effizient auf diesen zu suchen, sind im Bereich der Datenbanksysteme eine Reihe von Verfahren entwickelt worden. In den heutigen Datenbanksystemen werden hauptsächlich eindimensionale Zugriffsstrukturen, wie beispielsweise B-Bäume [BM 72], eingesetzt. Den effizienten Zugriff auf höherdimensionale Daten (d.h. Datensätze mit mehreren Schlüsselattributen) unterstützen mehrdimensionale Zugriffsstrukturen.

Die Implementierung des in diesem Artikel vorgestellten Feature-Index basiert auf dem R*-Baum. Der R*-Baum ist eine Indexstruktur, die zur effizienten Abspeicherung und Suche von räumlich ausgedehnten Objekten (z.B. Landkarten oder CAD-Bauteilen) in Datenbanksystemen entwickelt wurde [BKSS 90]. Die Idee des R*-Baums besteht darin, die abzuspeichernden Objekte durch k-dimensionale Rechtecke zu approximieren und den Raum dieser Rechtecke hierarchisch zu partitionieren. Die Partitionen sind selbst wieder k-dimensionale Rechtecke, die einander überlappen dürfen und den Datenraum nicht vollständig überdecken müssen. Diese Partitionierung wird mit Hilfe einer Baumstruktur dargestellt. Der Prozeß der Anfragebearbeitung läßt sich mit Hilfe eines R*-Baums auf die relevanten Ausschnitte des Datenraums beschränken. Abbildung 7 zeigt als Beispiel einen R*-Baum für zweidimensionale Rechtecke.

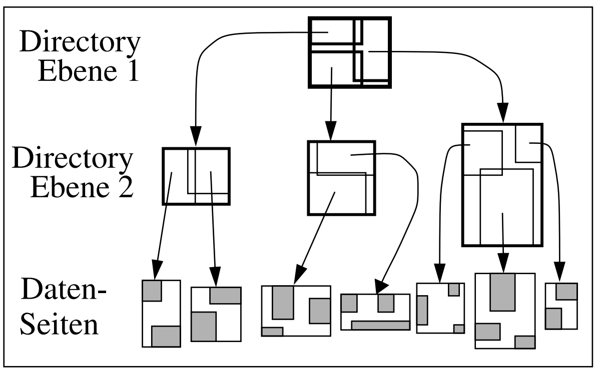

Abb. 7: Beispiel eines R*-Baums für zwei Dimensionen

Mit *Seite* bezeichnen wir diejenige Einheit des Plattenspeichers, die als Ganzes in den Hauptspeicher übertragen wird. Jedes Rechteck in einer Directory-Seite ist das *minimal umgebende achsenparallele Rechteck (MUR)* für alle Rechtecke in allen Directory- oder Datenseiten, die im zugehörigen Teilbaum liegen. Eine Seite besteht aus einer Menge von Einträgen. Jeder Eintrag in einer Directory-Seite besteht aus einem MUR und einem Verweis auf eine Seite (Directory- oder Datenseite), die den durch das MUR definierten Teil des Datenraums näher beschreibt. Ein Eintrag in einer Datenseite besteht aus einem MUR und einem Verweis auf die exakte Objekt-Repräsentation, falls die Objekte nicht rechteckig sind.

Durch Einfügen von neuen Objekten in Datenseiten bzw. durch Aufteilen von Sohnseiten kann eine Seite überlaufen. Der Aufteilung einer Menge von Rechtecken in zwei Mengen liegen mehrere Kriterien zugrunde: Die Überlappung sowie der Umfang der Directory-Seiten und der Fehlraum zwischen den Rechtecken soll minimiert werden.

Die Basis-Anfragen, die man mit Hilfe eines R*-Baums beantworten kann, sind die *Point-Queries* (liefern alle Objekte, die einen gegebenen Anfragepunkt enthalten) und die *Window-Queries* (liefern alle Objekte, die ein gegebenes rechteckiges Anfragefenster schneiden). Als Beispiel stellen wir den Algorithmus zur Beantwortung von Window-Queries vor, da die Anfragebearbeitung im Feature-Index darauf basiert. Wir rufen folgenden Algorithmus mit der Wurzel des R*-Baums (Page = Wurzel) und einem gegebenen Anfragefenster (Window) auf:

```
WindowQuery (Page, Window);
   FOR ALL Entry ∈ Page DO
      IF Window INTERSECTS Entry.Rectangle THEN
         IF Page = DataPage THEN
            Write (Entry.Rectangle)
         ELSE
            WindowQuery (Entry.Subtree^, Window).
```

Der obige Algorithmus durchsucht nur solche Rechtecke im R*-Baum nach Antworten, die einen nichtleeren Durchschnitt mit dem Anfrage-Window besitzen. Abbildung 8 veranschaulicht den Effizi-

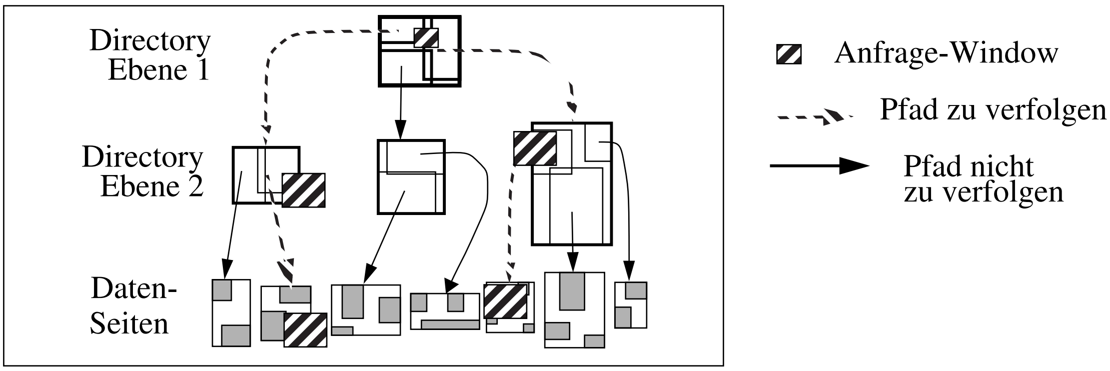

Abb. 8: Beantwortung einer Window-Query in dem Beispiel-R*-Baum

enzgewinn, den man durch die Reduktion der Zahl der nötigen Seitenzugriffe erhält.

Experimentelle Untersuchungen mit realistischen Testdaten [BKSS 90] zeigen, daß der R*-Baum eine ausgezeichnete Effizienz besitzt. Ein weiteres Ergebnis ist, daß der R*-Baum nicht nur für ausgedehnte Objekte, sondern genausogut auch für punktförmige Objekte geeignet ist. Die Leistungsfähigkeit der Anfragebearbeitung läßt sich durch Einsatz von Objekt-Approximationen und Objekt-Dekompositionen noch steigern [BKS 93][KSB 93].

7. Zusammenfassung und Ausblick

Bei der Suche nach geeigneten Dockingpartnern für ein gegebenes Molekül in einer Proteindatenbank muß eine große Menge von Proteinen durchsucht werden. Wegen ihres hohen Bedarfs an Rechenzeit kommen die meisten aus der Literatur bekannten Docking-Algorithmen, bei denen zwei Moleküle relativ zueinander rotiert, verschoben und bewertet werden, für eine solche Suche nicht in Betracht. Wir verfolgen deshalb in diesem Artikel den Ansatz, rotations- und translationsunabhängige Kennzahlen zu definieren. Die Suche nach komplementären Oberflächen-Features kann somit unabhängig von der Lage der einzelnen Proteine im Raum erfolgen.

Der vorgestellte Feature-Index ermöglicht eine effiziente Suche auf einer großen Datenbank von Proteinen. Die Algorithmen und Datenstrukturen des Feature-Index basieren auf denen des R*-Baums, der ein effizientes Verfahren zur Suche auf einer großen Menge von mehrdimensionalen ausgedehnten Raumobjekten darstellt.

Bei der Suche nach Oberflächen-Features muß berücksichtigt werden, daß zu einer Anfrage-Region nicht nur das exakte Komplement, sondern auch zum Komplement ähnliche Regionen in der Datenbank gefunden werden sollen. Im Feature-Index wird dies dadurch unterstützt, daß alle Regionen als Antwort geliefert werden, deren Feature-Mengen mit der Feature-Menge der Anfrage-Region hinreichend stark überlappen. Dadurch wird der Feature-Index robust.

Für die Suche nach Proteinen mit komplementären Features wird die Oberfläche des Anfrageproteins regionalisiert und die Kennzahlen werden komplementiert. Im Feature-Index wird dann nach Regionen mit Features gesucht, die zu denen des komplementierten Anfrageproteins ähnlich sind. Werden passende gefunden, so kann daraus eine erste Bewertung der Dockingstelle mit Hilfe des Ähnlichkeitsgrades vorgenommen werden. In den folgenden Schritten wird ein Docking-System dann die zu den Regionen gehörenden Moleküle für die gefundene relative Lage auf räumliche Überlappung überprüfen und ihre molekularen Wechselwirkungen berechnen.

Gemeinsam mit unseren Projektpartnern entwickeln wir derzeit zur Beschreibung von Oberflächenfeatures geeignete Kennzahlen und Verfahren zur Regionalisierung. Wir führen eine Evaluierung des Feature-Index mit Hilfe der SA- und STI-Kennzahlen durch. In Zukunft werden wir neben weiteren geometrischen zusätzlich auch physikochemische Kennzahlen (z.B. Potentialwerte) in den Index aufnehmen. Es stehen uns noch keine Methoden zur Verfügung, die für zwei Regionen sicher entscheiden können, ob sie docken. Wir benutzen daher im Moment zum Erzeugen von Test-Anfragen in der Proteindatenbank [Ber 77] enthaltene Komplexe, die durch Docking aus mehreren Proteinen entstanden sind.

Wir untersuchen u.a. folgende Fragen:

- Wie gut stimmen die durch die Regionalisierung erhaltenen Regionen mit den Flächen der bekannten Docking-Stellen überein? Kann man insbesondere annehmen, daß eine Docking-Stelle aus einer einzigen Region besteht, oder setzt sich eine Docking-Stelle aus mehreren Regionen zusammen?
- Werden vom Feature-Index alle aus der Proteindatenbank bekannten Docking-Stellen gefunden? Wie hoch ist der Anteil der Fehlantworten, d. h. der gelieferten Regionen, die sich nicht zum Docking eignen?

Erste Antworten auf diese Fragen werden zum Zeitpunkt der Tagung “Bioinformatik - Computereinsatz in den Biowissenschaften” vorliegen.

Danksagung

Wir danken Xiaowei Xu und unseren Kollegen im Verbundprojekt BIOWEPRO für fruchtbare Diskussionen, Thomas Schmidt für die Erstellung des Programms CAPS, sowie Thomas Reiter und Stefan Wirth für die Evaluierung des Feature-Index.

Literaturhinweise

[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.

[BKS 93] Brinkhoff T., Kriegel H.-P., Schneider R.: *'Comparison of Approximations of Complex Objects used for Approximation-based Query Processing in Spatial Database Systems'*, Proc. 9th Intl. Conf. on Data Engineering, Vienna, Austria, 1993.

[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.

[BM 72] Bayer R., McCreight E. M.: *'Organization and Maintenance of Large Ordererd Indexes'*, Acta Informatica, Vol. 1, No.1, 1972, pp. 173-189.

[CM 90] Colloc'h N., Mornon J.-P.: *'A new tool for the qualitative and quantitative analysis of protein surfaces using B-spline and density of surface neighborhood',* Journal of Molecular Graphics, Vol. 8, 1990, pp. 133-140.

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

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

[Con 92] Connolly M. L.: *'Shape Distribution of Protein Topography',* Biopolymers, Vol. 32, 1992, pp. 1215-1236.

[DO 93] Duncan B. S., Olson A. J.: *'Shape Analysis of Molecular Surfaces',* Biopolymers, Vol. 33, 1993, pp. 231-238.

[PBEÖ 90] Projektträger des BMFT für Biologie, Energie, Ökologie (Hrsg.): *'Zellen und Computer - Alternativen zum Tierversuch'*, Forschungszentrum Jülich, 1990.

[Hei 93] Heiden W.: *'Methoden zur computergestützten Untersuchung selektiver Oberflächeneigenschaften von Proteinen'*, Dissertation, TH Darmstadt 1993.

[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 U.S.A., Vol. 89, 1992, pp. 2195-2199.

[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.

[Len 93] Lengauer T.: *'Algorithmic Research Problems in Molecular Bioinformatics'*, Arbeitspapiere der GMD 748, May 1993.

[Nie 90] Niemann H.: *'Pattern Analysis and Understanding'*, 2nd ed., Springer 1990, pp. 103-110.

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

[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.

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

[WS 92] Walls P. H., Sternberg M. J.: *'New Algorithm to Model Protein-Protein Recognition Based on Surface Complementary: Applications to Antibody- Antigen Docking',* Journal of Molecular Biology, Vol. 228, 1992, pp. 277-297.

[ZHSB 92] Zachmann C.-D., Heiden W., Schlenkrich M., Brickmann J.: *'Topological Analysis of Complex Molecular Surfaces'*, Journal of Computational Chemistry, Vol. 13, No. 1, 1992, pp. 76-84.