Können Proteine im Computer docken?[1]

Hans-Peter Kriegel, Thomas Schmidt, Thomas Seidl

Proteine spielen eine grundlegende Rolle in allen Organismen: Sie sind die ausführenden Instanzen in den lebenswichtigen Prozessen der Verdauung, des Stoffwechsels und des Immunsystems. Die Funktion solcher Eiweißmoleküle besteht in der Wechselwirkung mit anderen Molekülen. Dabei docken die Moleküle aneinander an und bilden einen Komplex. Eine wichtige Grundregel der Interaktion ist das “Schlüssel-Schloß-Prinzip”: Die Oberflächen der dockenden Moleküle weisen eine komplementäre Geometrie auf.

Das Schlüssel-Schloß-Prinzip ist der Ausgangspunkt, um molekulare Funktionen weiter zu enträtseln. Ein erklärtes Ziel aktueller biomolekularer Forschung ist es, die Arbeitsweise von Molekülen nachzubilden und zu prognostizieren. Das Anwendungsspektrum solcher Simulationen erstreckt sich auf Molekularbiologie, Medizin und Pharmakologie. Man erhofft sich beispielsweise, sowohl Wirkungen von Medikamenten als auch ihre Nebenwirkungen bezüglich der Immunoglobuline eines Organismus vorhersagen zu können. Durch Computersimulation sollen aufwendige Laborexperimente und Tierversuche eingeschränkt werden. Ein großer Fortschritt wäre erreicht, wenn mit Hilfe solcher Simulationen zu einem gegebenen Protein ein oder mehrere Interaktionspartner aus einer großen Menge von Proteinen effizient ermittelt werden könnten.

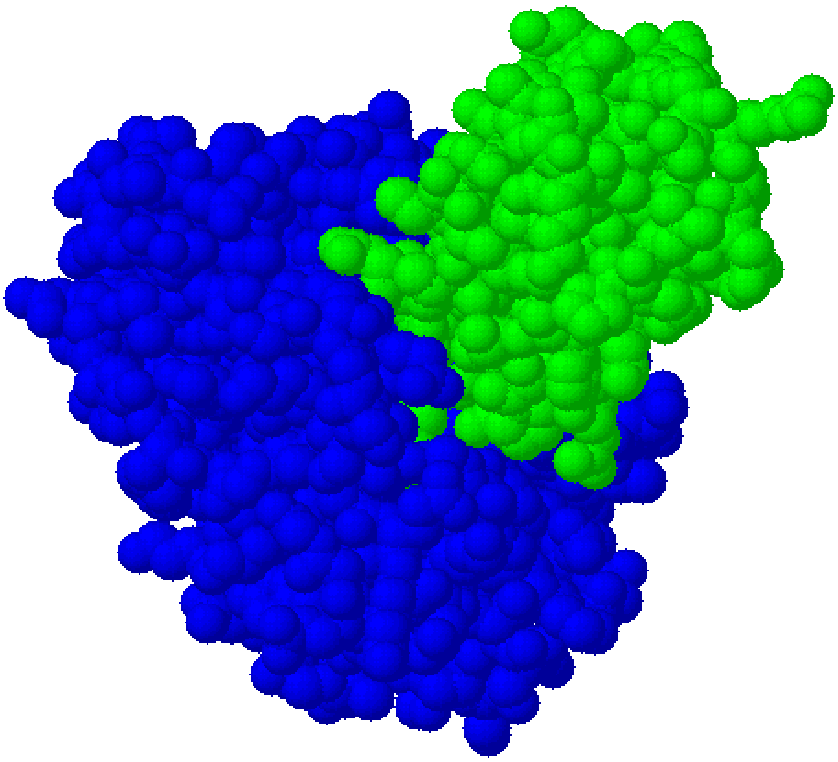


Abbildung 1: Komplex aus Trypsin und Trypsin-Inhibitor

In der Brookhaven Protein Data Bank (PDB) sind verschiedene Proteinkomplexe abgespeichert, für die man die Dockingkonstellation genau kennt. Anhand dieser Beispielkomplexe lassen sich Dockingalgorithmen überprüfen. Die Aufgabe besteht darin, zu einem Molekül den richtigen Dockingpartner in der Datenbank zu finden (1:n-Docking). Alleine für das 1:1-Docking gibt es schon sechs Freiheitsgrade, um ein Molekül an ein anderes anzulagern.

Im Rahmen des BMBF-Verbundprojekts BIOWEPRO (Biomolekulare Wechselwirkungen von Proteinen) wird am Institut für Informatik der LMU zusammen mit der Gesellschaft für Biotechnologische Forschung in Braunschweig, der Arbeitsgruppe Angewandte Informatik der Universität Bielefeld und dem Max-Planck-Institut für Biophysikalische Chemie in Göttin-

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

gen ein Proteindatenbanksystem entwickelt, das Biologen bei der Suche nach Dockingpartnern unterstützen soll [Sch 95] [Ald 94]. Für ein gegebenes Anfrageprotein soll das System Antworten auf folgende Fragen liefern:

- Welche Proteine in der Datenbank kommen als Dockingpartner in Frage?
- Auf welcher Region der Oberfläche findet die Komplexbildung jeweils statt?

Als Datengrundlage stehen die kristallographisch ermittelten Atomkoordinaten von Proteinen und Proteinkomplexen in der Brookhaven Protein Data Bank (PDB) zur Verfügung. Die PDB enthält derzeit knapp 3.000 Proteine, Enzyme und Viren und wächst ständig.

Die Vorhersage der relativen Position zweier Proteine in einem Komplex macht im Grunde eine Berechnung der freien Energie des Gesamtsystems von Bindungspartnern und umgebendem Lösungsmittel für alle möglichen Positionen erforderlich. Werden Positionen gefunden, in denen die freie Energie geringer ist, als die des Systems der einzelnen Moleküle, so bestimmt das Minimum unter ihnen die gesuchte Dockingposition. Da diese Berechnungen jedoch extrem aufwendig sind, ist es erforderlich, effizienter zu bestimmende Kriterien zur Lösung des Problems heranzuziehen. So müssen beim Docking zwei Bedingungen erfüllt sein: Zum einen das bereits erwähnte Schlüssel-Schloß-Prinzip, d.h. die Dockingpartner weisen komplementäre Oberflächenregionen auf. Zum anderen spielen physiko-chemische Eigenschaften eine wichtige Rolle, wie z.B. elektrostatische Wechselwirkungen oder die Möglichkeit Wasserstoffbrücken auszubilden, die den Komplex stabilisieren.

Es hat sich gezeigt, daß sich viele Dockingpositionen allein aufgrund der komplementären Oberflächengeometrie bestimmen lassen. Das Docking-Problem reduziert sich somit auf die Suche nach komplementären Oberflächenregionen. Leider ist dieses Problem selbst für zwei Proteine von immenser Komplexität. Die relative räumliche Position zweier Moleküle zueinander wird im folgenden als *Konstellation* bezeichnet. Der Raum der Konstellationen für zwei Proteine hat sechs Freiheitsgrade und ist unendlich groß: Das erste Protein kann im dreidimensionalen Raum entlang jeder der drei Raumachsen verschoben bzw. um jede dieser Achsen rotiert werden, um es an das zweite anzulagern.

Architektur der Anfragebearbeitung

In der Vergangenheit wurden bereits eine Reihe von Algorithmen zur Lösung des Docking-Problems auf Basis der Oberflächengeometrie vorgeschlagen. Alle bisherigen Verfahren konzentrieren sich auf das 1:1-Dokking, d.h. auf die Interaktion von zwei Proteinen. Diese Verfahren sind jedoch wenig geeignet, um Dockingpartner aus einer Datenbank mit einer großen Menge von Proteinen zu bestimmen. Denn dazu muß ein derartiges Verfahren auf alle Proteine in der Datenbank angewandt werden. Dies hätte einen Aufwand zur Folge, der linear mit der Anzahl der gespeichterten Proteine wächst. Um akzeptable Antwortzeiten zu erreichen, ist jedoch ein sublinearer Aufwand wünschenswert.

Die enorme Größe des Konstellationsraums legt eine Anfragebearbeitung in mehreren Schritten nahe. Dabei wird im ersten Schritt, dem sog. Filterschritt, die Menge der Docking-Kandidaten aufgrund von schnell zu berechnenden Kriterien so weit wie möglich eingeschränkt. Die verbleibenden Kandidaten werden anschließend in aufwendigeren Verfeinerungsschritten genauer untersucht. Dieses Vorgehen der mehrstufigen Anfragebearbeitung hat sich im Bereich der Geo-Datenbanken bewährt. Die einzelnen Schritte werden im folgenden kurz erläutert [EKSX 95]:

Im Filterschritt werden diejenigen Proteine aus der Datenbank bestimmt, die einen zum Anfrageprotein komplementären Oberflächenbereich aufweisen. Zu diesem Zweck werden die Proteine beim Speichern in die Datenbank auf folgende Weise vorverarbeitet: Zunächst wird ihre Oberfläche berechnet und daraus werden einzelne Segmente extrahiert, die als Anlagerungsstellen in Frage kommen. Diese Segmente, die jeweils durch eine Menge gleichmäßig verteilter Oberflächenpunkte beschrieben werden, werden in der Datenbank abgelegt. Bei der Anfragebearbeitung wird die Oberfläche des Anfrageproteins analog segmentiert. Für jedes der so erhaltenen Segmente wird in der Datenbank nach Segmenten gesucht die

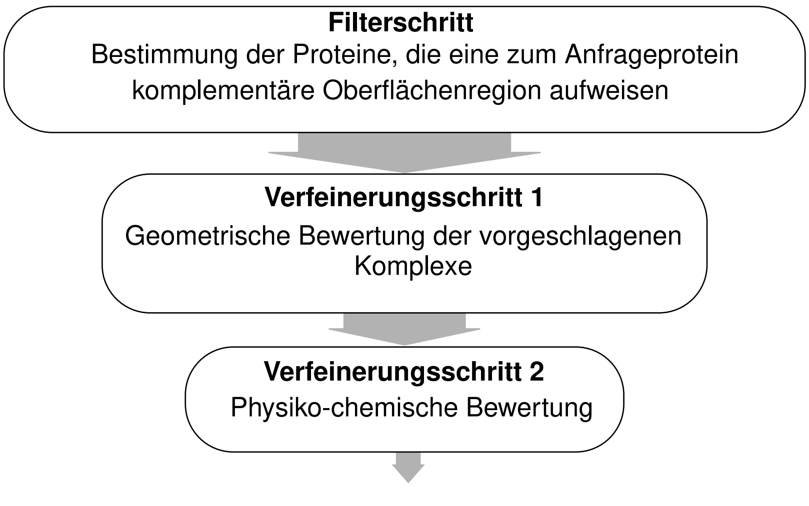


Abbildung 2: Ablauf der Anfragebearbeitung

Die Bearbeitung einer Docking-Anfrage in dem Datenbanksystem läuft in mehreren Schritten ab. Im Filterschritt werden aus der großen Proteinmenge diejenigen Kandidaten bestimmt, die ein zum Anfrageprotein "passendes" Oberflächensegment aufweisen. Durch den Einsatz von Indexstrukturen müssen dazu nicht alle Proteine untersucht werden. Die so erhaltene Kandidatenmenge wird in den Verfeinerungsschritten auf eine sehr kleine Zahl weiter reduziert.

ein gewisses Maß an Komplementarität aufweisen. Durch den Einsatz von mehrdimensionalen Indexstrukturen müssen dazu nicht alle Proteine untersucht werden. Im weiteren Verlauf der Anfragebearbeitung müssen nur noch solche Proteine betrachtet werden, von denen mindestens ein passendes Segment im Filterschritt gefunden wurde.

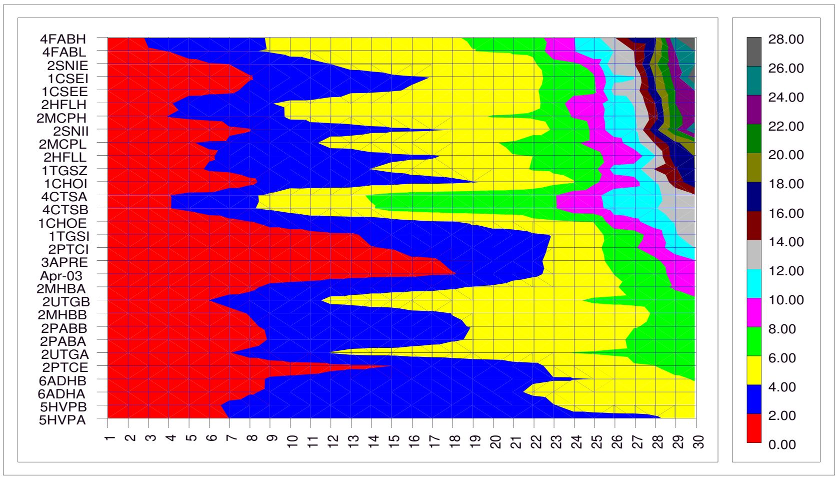


Abbildung 4: Werteverteiluung des Ähnlichkeitsmaßes

Für die Beispielsmenge von 30 Dockingsegmenten erkennt man, daß sich das vorgeschlagene Ähnlichkeitsmaß für eine Unterscheidung der Formähnlichkeit eignet. Am linken Rand sind die Moleküle aufgeführt, deren Dockingstellen untersucht wurden. Nach rechts sind die Ähnlichkeitswerte aufsteigend als Farbwerte aufgetragen.

Aus jedem der Antwortsegmente wird durch Übereinanderlegen mit dem korrespondierenden Segment des Anfrageproteins eine Konstellation berechnet. Diese wird in den Verfeinerungsschritten bewertet. Die Bewertung kann zum einen unter geometrischen Gesichtspunkten erfolgen und zum anderen unter physiko-chemischen Aspekten. Aus geometrischer Sicht können alle Konstellationen verworfen werden, bei denen sich die beiden Moleküle zu stark überschneiden oder die Kontaktfläche nicht groß genug ist. Die übriggebliebenen Konstellationen werden einer Energiebewertung unterzogen. Auf diese Weise läßt sich die Zahl der möglichen Dockingkandidaten und -konstellationen auf eine sehr kleine Zahl reduzieren. Endgültige Gewißheit über die Interaktion zweier Proteine kann aber letztlich nur ein Laborexperiment geben. Der Ablauf der Anfragebearbeitung ist in Abbildung 2 noch einmal zusammengefasst.

Der Filterschritt ist aus mehreren Gründen von zentraler Bedeutung für das Verfahren und soll daher im folgenden näher beleuchtet werden. Er muß einerseits so selektiv sein, daß in den anschließenden, aufwendigen Verfeinerungsschritten nur eine kleine Anzahl von Kandidaten überprüft werden muß. Andererseits darf er keine wirklichen Dockingkandidaten fälschlicherweise ausschließen. Es ergeben sich folgende Anforderungen an das Datenbanksystem: Oberflächensegmente, die als Mengen von 3D-Punkten vorliegen, müssen so gespeichert werden, daß eine Suche nach komplementären Formen möglich ist. Dabei darf die Lage und Orientierung der Segmente im Raum keine Rolle spielen, d. h. es muß Translations- und Rotationsinvarianz gewährleistet sein. Im folgenden werden zwei Ansätze dazu vorgestellt.

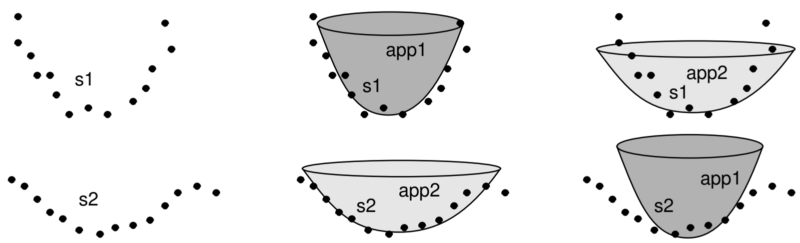


Abbildung 3: Ähnlichkeit über geometrische Approximation

Die Ähnlichkeit zweier Segmente s_1 und s_2 läßt sich über geometrische Approximationen bestimmen. Dazu wird ein Approximationsmodell ausgewählt, im Beispiel ein Paraboloid. Für beide Segmente werden die Approximation app_1 und app_2 mit dem geringsten Approximationsfehler nach der "Methode der kleinsten Fehlerquadrate" berechnet. Um den Abstand der beiden Segmente s_1 und s_2 zu quantifizieren, vergleicht man nun ein Segment mit der Approximation des anderen Segmentes. Die Summe der dadurch ermittelten wechselseitigen Approximationsfehler liefert – nach einem Normierungsschritt – das gewünschte Abstandsmaß $d(s_1, s_2)$.

Geometrische Approximation

Ein Ansatz, um geometrisch komplementäre oder ähnliche Oberflächensegmente zu bestimmen, besteht in der Verwendung von Approximationen durch einfache geometrische Modelle. In Betracht kommen dafür Paraboloide verschiedenen Grades, trigonometrische Polynome oder auch andere Flächenfunktionen. Die Ähnlichkeit von Segmenten läßt sich dann anhand der wechselseitigen Approximationsfehler bestimmen.

In Abbildung 4 ist dargestellt, wie sich dieses Ähnlichkeitsmaß für eine kleine Menge von 30 Dockingsegmenten verhält. Als Approximationsmodell wurde ein Paraboloid zweiten Grades gewählt.

Geometrische Formkennzahlen

Ein alternativer Ansatz zur geometrischen Approximation 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 Dokkingsuche komplementäre Segmente gesucht werden, müssen sich die Kennzahlen auch leicht komplementieren lassen. Für unsere Experimente haben wir *Solid Angle* und *Local Shape Index* ausgewählt (siehe Abbildung 5).

Um mit Hilfe der Kennzahlen ein Oberflächensegment zu charakterisieren, berechnen wir zunächst für alle Punkte des Segmentes deren Kennzahlen. Diese werden dann zusammengefaßt in einem Intervall dargestellt. Da wir mehrere Kennzahlen gleichzeitig verwenden, entsteht dabei ein Quader im Raum der Kennzahlen.

Zur Bearbeitung einer Dockinganfrage schließlich werden die Oberflächensegmente des Anfragemoleküls der gleichen Kennzahlenberechnung unterworfen. Nach der Komplementierung entsteht nun analog ein Kennzahlenquader. Dieser dient dazu, in effizienter Weise diejenigen Segmente in der Datenbank zu ermitteln, die zu den Anfragesegmenten komplementär sind.

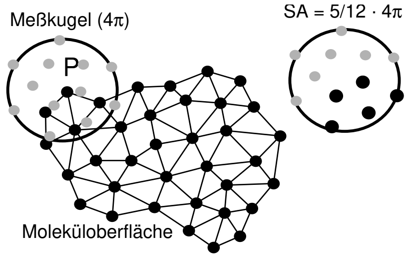
Ausblick

Die Bearbeitung von Dockinganfragen ist eingebettet in ein objektorientiertes Datenbanksystem, das sowohl die chemische als auch die räumliche Struktur der Moleküle abspeichert. Dadurch stehen Informationen bereit, um auch physiko-chemische Eigenschaften wie Hydrophobizität oder elektrostatisches Potential an der Moleküloberfläche in den Filterschritt der Dockingsuche einzubeziehen [KSS 95] [SK 95].

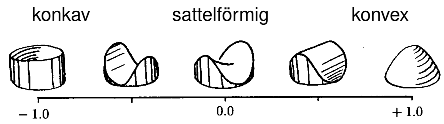
Die vorgestellten Verfahren werden derzeit auf einer Auswahl von mehreren hundert Molekülen getestet. Das erste Ziel dabei ist, die bekannten Komplexe der PDB zu rekonstruieren. Für eine Auswahl von 20 Komplexen wurden bereits sehr zufriedenstellende Ergebnisse erzielt [Sch 95]. Mit dem dabei gewonnen Wissen über die bekannten Interaktionen werden dann die Verfahren so verbessert und abgestimmt, daß auch neue Komplexe vorhergesagt werden können.

Abbildung 5: Geometrische Formkennzahlen

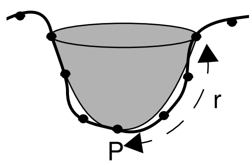
Die geometrischen Kennzahlen beschreiben die Form der Oberfläche in der Umgebung eines Punktes P. Über den Radius hat man jeweils eine Skalierungsmöglichkeit, wie weit diese Umgebung reichen soll. Als geeignet haben sich Werte zwischen 3 Å und 8 Å erwiesen.



Der *Solid Angle* (SA) beschreibt den Öffnungswinkel der Oberfläche um P. Bei der zählenden Ermittlung des SA legt man eine Meßkugel M zugrunde, auf deren Oberfläche eine bestimmte Anzahl von Punkten gleichmäßig verteilt ist. Man legt nun die Kugel M um den Punkt P und zählt diejenigen Punkte auf M, die nicht im Inneren des Moleküls liegen. Nach einer Normierung erhält man den Solid Angle im Punkt P. Den komplementären Wert des SA erhält man durch Ergänzung auf 4π.



Der *Local Shape Index* (LSI) beschreibt, welcher Formtyp an der Oberfläche um den Punkt P vorliegt. Er liefert eine Charakterisierung für konkave, konvexe und sattelförmige Flächenstücke. Den komplementären Wert des LSI erhält man durch Wechsel des Vorzeichens.



Zur Berechnung des LSI am Punkt P wird ein elliptisches bzw. hyperbolisches Paraboloid in die Umgebung von P in die Oberfläche eingepaßt. Aus den beiden Hauptkrümmungen k_{min} und k_{max} des Paraboloids berechnet man den LSI als $2/\pi \cdot \mathrm{atan}(k_{max}+k_{min})/(k_{max}-k_{min})$ **.**

Prof. Dr. Hans-Peter Kriegel
Thomas Schmidt
Thomas Seidl
Lehr- und Forschungseinheit für Datenbanksysteme
Institut für Informatik, LMU
Oettingenstr. 67, 80538 München
http://www.dbs.informatik.uni-muenchen.de

LITERATUR

(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.

(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, pp. 373-382.

(KSS 95) Kriegel H.-P., Schmidt T., Seidl T.: 'Moleküloberflächen in einer objektorientierten Proteindatenbank', 3. GI-Fachtagung "Bioinformatik – Informatik in den Biowissenschaften, Braunschweig, 9.-11. Okt. 1995, erweiterte Zusammenfassung.

(Sch 95) Schomburg D., et al.: 'BIOWEPRO – Biomolekulare Wechselwirkungen von Proteinen', in BMBF (Hrsg.): Tagungsband BMBF-Statusseminar Bioinformatik "Molekulare Bioinformatik und Evolutionäre Algorithmen", Braunschweig, 7.-8. Nov. 1995.

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