3. GI-Fachtagung “Bioinformatik — Informatik in den Biowissenschaften”, Braunschweig, 09.-11.10.95

Moleküloberflächen in einer objektorientierten Proteindatenbank[1]

Hans-Peter Kriegel, Thomas Schmidt, Thomas Seidl

Institut für Informatik, Universität München
Leopoldstr. 11B, D-80802 München
{ kriegel | schmidtt | seidl }@informatik.uni-muenchen.de

Schlüsselwörter: Datenbanken in der Molekularbiologie, 3D-Strukturdaten, Modellierung und Speicherung von Connolly-Oberflächen, Mehrfachrepräsentation

Für die Untersuchung der Funktionsweise biologischer Makromoleküle ist die Kenntnis von Strukturdaten eine wichtige Voraussetzung. So ist die 3D-Struktur von Proteinen die Grundlage für die Bestimmung möglicher Dockingpartner. Im Rahmen des Verbundprojekts BIOWEPRO (Biomolekulare Wechselwirkugen von Proteinen) wird ein Protein-Docking-Datenbanksystem entwickelt, das Biologen bei der Suche nach Dockingpartnern unterstützen soll [Ald 94].

Die dabei nötigen Strukturdaten sind in der Brookhaven Protein Data Bank (PDB) als Textfiles verfügbar [PDB 95]. Für das Docking ist die alleinige Darstellung der Proteinstruktur durch die Atomkoordinaten jedoch nicht ausreichend; vielmehr ist die Repräsentation der Moleküloberfläche als Zwischenstufe für nachfolgende Formbeschreibungen notwendig. In der Vergangenheit wurden auf der Grundlage des PDB-Formats verschiedene Schemata zur Speicherung von Proteinstrukturdaten vorgeschlagen, die auf unterschiedlichen Datenmodellen basieren. Da jedoch keiner dieser Ansätze die Darstellung von Moleküloberflächen berücksichtigt, wurde ein Schema entwickelt, das diesem Gesichtspunkt Rechnung trägt [Som 94].

Bei der Auswahl des Datenmodells war zu berücksichtigen, daß Proteine komplex aufgebaute, hierarchisch strukturierte Objekte sind. Für derartige Daten ist das relationale Modell wenig geeignet, da es lediglich skalare Wertebereiche für Attribute zuläßt. Dagegen ist für die Darstellung der Oberflächen in einer Datenbank ein Datenmodell vorteilhaft, das z.B. die Definition abstrakter Datentypen ermöglicht, um so von den Implementierungsgesichtspunkten der Oberfläche zu abstrahieren. Dadurch kann dem Benutzer eine wesentlich einfachere Schnittstelle zum Zugriff auf die Oberfläche angeboten werden. Aus diesem Grunde fiel die Wahl auf ein objektorientiertes Datenmodell. Die Implementierung erfolgte mit dem kommerziellen objektorientierten Datenbanksystem ObjectStore [OHMS 92].

In der nachfolgenden Abbildung ist das Schema für die Proteinoberflächen in unserer Datenbank illustriert. Eine Klasse ist dabei durch ein Rechteck dargestellt, ihre Methoden werden durch die Beschriftung der von ihr ausgehenden Pfeile bezeichnet. Diese verweisen auf den Ergebnistyp der jeweiligen Methode. Gestrichelte Pfeile kennzeichnen isa-Beziehungen.

Als Oberflächenbegriff für Moleküle verwenden wir die Connolly-Oberfläche, die aus konvexen, sattelförmigen und konkaven Flächenstücken (Patches) besteht [Con 83]. 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, da die hohe Regularität der Oberflächenstruktur ausgenutzt wird. Andererseits stehen eine Reihe von Operationen zur Verfügung, die ein Navigieren auf der Oberfläche ermöglichen.

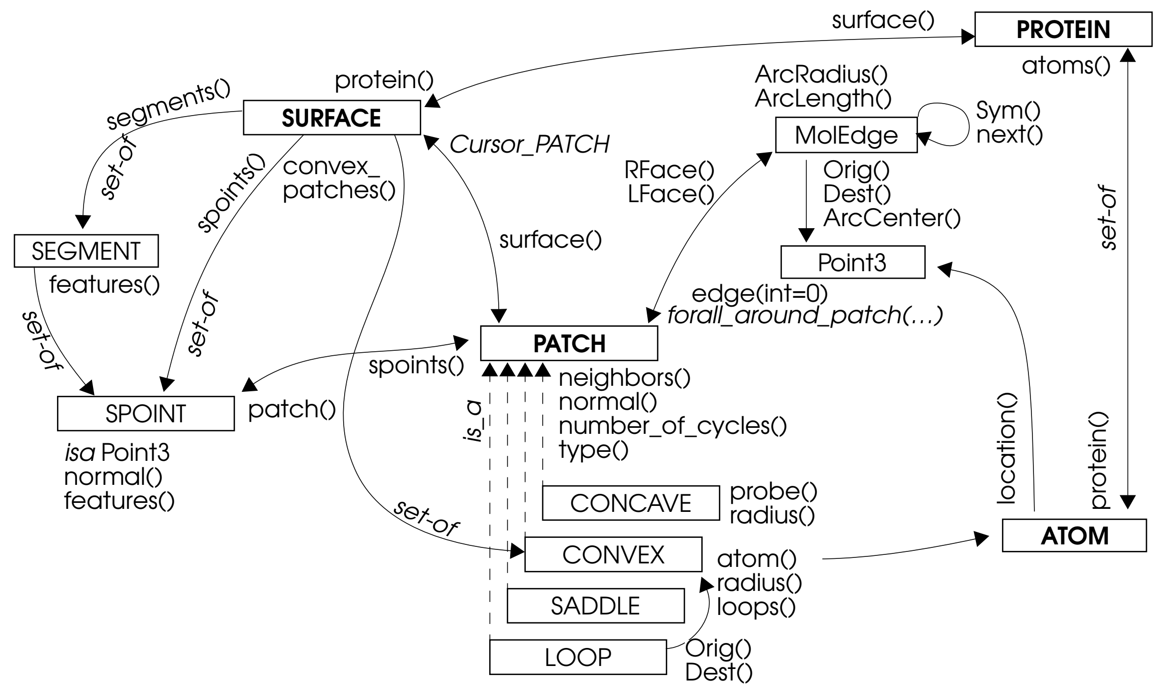
Die analytische Repräsentation durch Patches bringt zudem den Vorteil der leichten Ableitbarkeit anderer Oberflächendarstellungen. So bietet unsere Speicherungsstruktur neben der Connolly-Oberfläche auch den Zugriff auf die zugängliche Oberfläche [Ric 77], sowie die Möglichkeit, Punkte auf der Oberfläche 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

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.

begrenzen, erhält man auf einfache Weise ein Kanten- und Flächenmodell der Molekühloberfläche, das durch allgemeine Visualisierungstools graphisch dargestellt werden kann.

Auch in Bezug auf das Dockingproblem bieten sich eine Reihe von Anwendungen. Insbesondere läßt sich 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. Zur Ermittlung von Formbeschreibungen werden Moleküloberflächen lokal approximiert, indem man z.B. in die jeweiligen Regionen ein Paraboloid einpaßt.

Insgesamt bietet unser objektorientiertes Proteindatenbanksystem eine integrierte Speicherung der Strukturdaten von Proteinen, wie sie in der PDB verfügbar sind, zusammen mit den zugehörigen Oberflächen. Es stellt damit eine Basis für Aufbereitungen und Analysen dar, mit deren Hilfe Eigenschaften von Moleküloberflächen bestimmt werden können, die z.B. für das Docking von Bedeutung sind.



Schema für Proteinoberflächen

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.

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

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

[PDB 95] Protein Data Bank: *'Quarterly Newsletter No. 71 (Jan. 95)',* Brookhaven National Laboratory, Upton, NY, 1995.

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

[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), AAAI Press, 1995, pp. 350-358.

[SK 95b] 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; LNCS, Springer, 1995.

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