

Dissertation

Molecular Surface Comparison. A Versatile Drug Discovery Tool

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften unter der Leitung von

Prof. Dr. Hans Lohninger

164

Institut für Chemische Technologien und Analytik

eingereicht an der Technischen Universität Wien
Technisch-Naturwissenschaftliche Fakultät

von

Dipl.-Ing. Christian Hofbauer

Matr. Nr. 94 09 824

Gentzgasse 15/II/28

A-1180 WIEN

Wien, am 31. August 2004

Gewidmet dem Andenken an
meinen Großvater Josef Seidl

Zusammenfassung

Im modernen Entwicklungsprozess für neue Medikamente spielt der Einsatz von theoretischen und graphischen Methoden eine immer bedeutendere Rolle. Vor allem bei der Suche nach potentiellen Wirkstoffen für spezifische Proteine ist eine umfassende Kenntnis der Struktur des Proteinrezeptors von entscheidender Bedeutung. Neben der dreidimensionalen atomaren Zusammensetzung, wie sie von Röntgenstrukturanalyse und NMR Spektroskopie ermittelt werden kann, sind Oberflächendarstellungen besonders dazu geeignet, die Form und räumliche Ausdehnung eines Moleküls wiederzugeben. Zusätzlich lässt sich die Verteilung verschiedener chemischer und physikalischer Eigenschaften auf Moleküloberflächen sehr intuitiv darstellen und für optische Vergleiche verschiedener Strukturen einsetzen. Um die Analyse großer Substanzdatenbanken zu erleichtern, ist es allerdings notwendig die Suche nach ähnlichen Motiven auf Moleküloberflächen zu automatisieren. In der vorliegenden Arbeit wird das Computerprogramm SURFCOMP vorgestellt, das in der Lage ist, die Oberflächen verschiedener chemischer Verbindungen miteinander zu vergleichen und die gemeinsamen oder auch unterschiedlichen Eigenschaften zu ermitteln.

Da die Anzahl der Elemente, aus denen sich die Oberfläche eines Moleküls zusammensetzt, um ein Vielfaches höher ist als die Anzahl seiner Atome und Bindungen, ist es notwendig eine Darstellung zu wählen, die mit wenigen Elementen die charakteristischen Eigenschaften der Moleküloberfläche ausreichend beschreibt. In SURFCOMP werden dazu alle kritischen Punkte auf den beiden Oberflächen verwendet. Diese liegen entweder an der Spitze einer konvexen Region (Hügel) oder am Boden einer konkaven Region (Tal). Mit diesen Punkten wird nun ein Assoziationsgraph gebildet, der alle potentiell ähnlichen kritischen Punktpaare beider Oberflächen enthält. Dieser wird mit Hilfe von mehreren Filtern, die jene Paare eliminieren, die entweder aus chemischer oder geometrischer Sicht nicht zusammenpassen, soweit vereinfacht, dass gemeinsame Motive mit Hilfe einer Cliquesuche erkannt werden können. Die dabei detektierten Ähnlichkeiten sind überwiegend lokaler Natur. Um ein umfassendes Bild aller möglichen Gemeinsamkeiten auf beiden Oberflächen zu erhalten wird abschließend eine hierarchische Clusteranalyse aller gefundenen lokalen Ähnlichkeiten durchgeführt.

Mit dem vorliegenden Programm konnte zunächst die relative Orientierung von acht verschiedenen Inhibitoren im Rezeptor von Thermolysin erfolgreich reproduziert werden, was durch den Vergleich mit bereits publizierten Programmen belegt wurde. In weiterer Folge wurden die Auswirkungen von unterschiedlichen Algorithmen zur Generierung von Moleküloberflächen auf die Ergebnisse untersucht und die Flexibilität der hier vorgestellten Methode auf Konformationsänderungen der Moleküle getestet. Schlussendlich konnte SURFCOMP erfolgreich zum Vergleich von Proteinoberflächen eingesetzt werden. Bei einer Gegenüberstellung zweier SH2 Domänen (SAP und EAT-2), die beide an dasselbe Signalpeptid gebunden waren, konnte eine Reihe von oberflächlichen Unterschieden auf Differenzen in der Aminosäuresequenz zurückgeführt werden. Der Nachweis von ähnlichen Motiven auf den Oberflächen der reaktiven Zentren von SAP und der Phosphatase PTP1B konnte die in biologischen Experimenten entdeckte Aktivität von SAP zur Dephosphorylierung von Phosphotyrosin bestätigen.

Abstract

Analysis of the distributions of physicochemical properties mapped onto molecular surfaces can highlight important similarities or differences between compound classes, contributing to rational drug design efforts [131]. This thesis will present a method that uses a combination of graph theory, computer vision and computational chemistry to detect local surface similarities between small and medium sized molecules. The present approach is based on 3D structure search where maximal common subgraph isomorphism is used to detect local similarities between the pharmacophoric feature points of different molecules [91]. The extension of this principle to molecular surfaces is cumbersome, because treatment of the complete set of surface points instead of just a few feature points with NP-hard graph algorithms is not feasible. In order to perform a reliable and fast detection of local surface similarities it is necessary to reduce the complexity of the problem by a set of filters that implement various geometric and physico-chemical heuristics.

To achieve this, a simplified representation of the surfaces is generated first consisting only of a set of critical points (corresponding to “hills” and “valleys” on the surface), augmented by their surrounding surface patches. Among all possible point pairs those are selected first that show sufficient chemical similarity, judged by means of a fuzzy dissimilarity index [48] between physicochemical properties mapped onto the surface points. Then the curvature patterns around all remaining point pairs are compared by harmonic shape image matching [145] to discard points that are not embedded in a similar shape. Finally the distances and angles between combinations of similar pairs are checked to be within certain boundaries to form an association graph that is simple enough for the clique detection. The cliques represent the local surface similarities and an alignment between the two molecular surfaces can be calculated based on the corresponding points. Finally the alignments can be clustered to reveal a picture of the total surface similarity between the two molecules.

The method was tested with a dataset of eight thermolysin inhibitors and recovered the correct alignments of the compounds bound in the active sites. The results were in good agreement with another surface-based comparison carried out on the same dataset [37]. Furthermore SURFCOMP was successfully applied to the comparison of protein active sites by means of spherical site selection and a scoring scheme that allows a fast identification of similar surface regions. A similarity search between the binding area of two similar SH2 domains (SAP and EAT-2) revealed interesting differences between their molecular surfaces, which could be assigned to the corresponding structural differences. Finally the surface similarities between SAP and tyrosine phosphatase PTP1B, which have been detected by SURFCOMP, support the idea that biological functions are strongly related to surface features, since SAP and PTP1B do not show any significant structural similarity.

Acknowledgements

This project would have never been successful without the support of my direct supervisor Dr. András Aszódi, to whom I want to express special thanks. His door and mind were always open for the discussion of emerging problems and new ideas and he managed to control the progress of my work while providing me with the utmost possible freedom. Besides András I want to thank my academic supervisor Prof. Dr. Hans Lohninger for his support of the project and his contribution to the publication of the method.

I would like to thank all my present and former colleagues from the In Silico Sciences Group at the Novartis Institute for Biomedical Research in Vienna – including Benjamin Almeida from the IT department – who I enjoyed working with and who helped me in many fruitful discussions that fostered the success of my project. I also want to acknowledge the help of Dr. Tamas Schweighoffer who provided me with insight into the biological problems related to the SH2 and phosphatase experiments. In addition I want to thank Drs Adrienne James and Torsten Schindler for revising parts of the manuscripts of the thesis and of the submitted publications.

Finally I want to thank my family for their unbroken support and love.

Table of Contents

1.	Introduction	1
1.1.	Previous work	2
1.2.	Concept.....	4
2.	Theory	5
2.1.	Molecular Surfaces	5
2.1.1.	Background.....	5
2.1.2.	Van der Waals Surface	5
2.1.3.	Isodensity Surface	6
2.1.4.	Solvent-Excluded and Solvent-Accessible Surfaces	8
2.1.5.	Representation of Molecular Surfaces	9
2.2.	Molecular Surface Properties.....	10
2.2.1.	Molecular Potentials	11
2.2.2.	Atomic Properties	12
2.2.3.	Surface Characteristics	12
2.3.	Feature Radius and Auto Correlation.....	14
2.4.	Fuzzy Logic	15
2.5.	Harmonic Images.....	17
2.5.1.	Concept.....	17
2.5.2.	Border Mapping	17
2.5.3.	Interior Mapping	18
2.5.4.	Generation and Comparison of Harmonic Shape Images	20
2.6.	Maximum Common Subgraph Isomorphism	23
3.	Methods	25
3.1.	General Concept	26
3.2.	Definition of Critical Points.....	28
3.3.	The Association Graph.....	29
3.4.	Generation of Surface Patches	29
3.5.	Fuzzy Filter	31
3.6.	Harmonic Shape Image Filter	31
3.7.	Distance Filter	33
3.8.	Overlap Filter	34
3.9.	Clique Detection and Clustering.....	35
3.10.	Scoring and Ranking	35
3.11.	Treatment of Protein Surfaces	37
3.12.	Implementation Details	38
3.12.1.	Software	38

3.12.2. Hardware and Computation.....	38
4. Computations and Results	40
4.1. Ligand Surfaces	40
4.1.1. Preparation of the Input Structures and Experimental Design	40
4.1.2. SURFCOMP Validation: Comparison of 8 Thermolysin Inhibitors	40
4.1.3. Ranking of Surface Alignments.....	47
4.1.4. Evaluation of Different Surface Types: Comparing DHFR ligands.....	55
4.1.5. Testing different conformations.....	61
4.2. Comparing Proteins: Surface Differences between SAP and EAT-2	67
4.2.1. Surface Comparison.....	70
4.2.2. Structural Investigations	72
4.3. Elucidating the Phosphatase Activity of SAP.....	77
5. Conclusion and Outlook	82
5.1. The Advantages of SURFCOMP	82
5.2. Discussion	83
5.3. Outlook.....	85
A The SURFCOMP Program Suite	87
A.1 Requirements	87
A.2 Availability.....	87
B Publications	88
C Abbreviations	90
D Indices	92
D.1 Figure Index	92
D.2 Table Index	93
D.3 Chart Index.....	93
E References	94

“Research is endlessly seductive, but writing is hard work”

Barbara Tuchman, *The Guns of August*.