Anhang D

Grundlagen der 3D-QSAR Methoden

In dieser Arbeit wurden zur Erstellung quantitativer 3D-Struktur-Aktivitäts-Modelle (3D-QSAR) die *Comparative-Molecular-Field-Analyse* (CoMFA) und die *Comparative-Binding-Energy-Analyse* (COMBINE) eingesetzt. Den Methoden liegt die Annahme zugrunde, dass ein quantitativer Zusammenhang zwischen den Unterschieden der Strukturen (hier WW-Domänen) und den Unterschieden der entsprechenden Aktivitäten (hier Bindungsaffinitäten) besteht. Für die Modellierung dieses Zusammenhangs wird ein statistisches Modell auf der Basis einer Reihe von Strukturen mit bekannter Aktivität (Trainingsdatensatz) trainiert.

Im ersten Schritt des Trainings wird jede der 1…*n* Strukturen durch je 1…*p* quantitative Deskriptoren beschrieben. Für den gesamten Trainingsdatensatz ergibt sich damit eine (*n,p*)-Deskriptoren-Matrix *X* mit *n* Fällen (Zeilen) und je *p* Deskriptoren (Spalten). Die beiden Methoden unterscheiden sich dabei nur in der Art der Deskriptoren. Während die CoMFA-Methode die Strukturen der ungebundenen Domänen durch ihre nicht-kovalenten molekularen Felder beschreibt (siehe D.1), werden durch die COMBINE-Methode die Domänen/Liganden-Komplexe durch alle paarweisen Interaktionsenergien zwischen den Aminosäureresten des Liganden und denen der Domäne beschrieben (siehe D.2). Da sich dabei eine große Anzahl an Deskriptoren pro Struktur – meist mehrere tausend – ergibt, entsteht dadurch eine „kurze" aber „breite" Matrix ($n << p$) der Deskriptoren.

Im zweiten Schritt des Trainings wird ein multivariates lineares statistisches Modell (Gleichung D:1) der Beziehung zwischen diesen Deskriptoren *X* (unabhängige Variablen) und den Aktivitäten *Y* (abhängige Variablen) berechnet. Gewöhnliche werden beide Matrizen zuvor zentriert (Mittelwert = 0) und skaliert (Standardabweichung = 1).

$$Y = XB + E \qquad (D:1)$$

3D-QSAR-Modell der Beziehung zwischen Struktur und Aktivität

Y ist die (*n,m*)-Aktivitäten-Matrix mit *n* Beobachtungen (Strukturen) und *m* Aktivitäten pro Struktur.
X ist die (*n,p*)-Deskriptoren-Matrix mit *n* Beobachtungen (Strukt.) und *p* Deskriptoren pro Struktur.
B ist die (*p,m*)-Regressions-Koeffizienten-Matrix mit *p* Koeffizienten pro *m* Aktivitäten.
E ist die (*n,m*)-Residual-Matrix, die nicht erklärte Fehlerstreuung.

Für die Berechnung des Modells wird die *Partial-Least-Squares*-Regression (PLS, siehe D.4) und nicht multiple lineare Regression verwendet, da die Anzahl der Deskriptoren *p* in der

Regel weitaus größer ist als die Anzahl der Strukturen *n* und die Deskriptoren meist kollinear, „verrauscht" sowie oft nicht normalverteilt sind. Sind keine Aktivitäten bekannt, können mit Hilfe der Hauptkomponenten-Analyse (PCA, siehe D.3) systematische Variationen in der Deskriptoren-Matrix X identifiziert werden, die eine Aussage über charakteristische Eigenschaften der analysierten Strukturen ermöglichen (siehe 3.3.3.1).

D.1 *Comparative-Molecular-Field-Analyse* (CoMFA)

Für die CoMFA-Methode ist die korrekte Superposition (Überlagerung) aller zu analysierenden 3D-Strukturen eine Grundvoraussetzung. Die überlagerten Strukturen werden danach in ein regelmäßiges 3D-Gitter eingebettet. Jede Domäne wird im Rahmen der CoMFA-Methode durch zwei nicht-kovalente molekulare Felder beschrieben. Diese repräsentieren die sterischen (Lennard-Jones-Potential) bzw. elektrostatischen Eigenschaften (Coulomb-Potential) der Domänen an definierten Punkten des regelmäßigen 3D-Gitters. Sie werden an Hand der Stärke der Wechselwirkungen einer Sonde (CH_3 bzw. H^+) an den Gitterpunkten mit den zu untersuchenden Domänen mittels eines Kraftfelds bestimmt. Von entscheidender Bedeutung für die Vergleichbarkeit der Felder unterschiedlicher Domänen ist dabei die korrekte Überlagerung der Strukturen.

D.2 Comparative-Binding-Energy-Analyse (COMBINE)

Im Rahmen der COMBINE-Analyse werden nicht nur die Strukturen der Domänen, sondern die Domänen/Liganden-Komplexe analysiert. Dazu werden die van-der-Waals (Lennard-Jones-Potential) und die elektrostatischen (Coulomb-Potential) Interaktionsenergien zwischen allen Aminosäureresten der Domäne und allen Resten des Liganden mittels eines Kraftfelds bestimmt. Von entscheidender Bedeutung für die Vergleichbarkeit der Interaktionsenergieterme unterschiedlicher Komplexe ist hierbei die strukturelle Homologie der Aminosäurepositionen in den verschiedenen Komplexstrukturen.

D.3 Hauptkomponenten-Analyse (PCA)

Mit Hilfe der Hauptkomponenten-Analyse (PCA) werden aus der Vielzahl der p Deskriptoren c Hauptkomponenten (latente Variablen/Faktoren) extrahiert (Gleichung D:2), welche für die Deskriptoren bestimmend sind (mit $c << p$). Geometrisch entspricht die PCA einer Projektion der durch X definierten Punktewolke aus einem p-dimensionalen kartesischen Koordinatensystem in ein neues c-dimensionales Koordinatensystem, wobei die Achsen den Hauptkomponenten entsprechen (Gleichung D:2).

$$X = TP + F \qquad (D:2)$$

Hauptkomponenten-Analyse (PCA) der Deskriptoren-Matrix X

X ist die (n,p)-Deskriptoren-Matrix mit n Beobachtungen und p Deskriptoren pro Struktur.
T ist die (n,c)-Faktorwert-Matrix mit c Faktorwerten für n Strukturen.
P ist die (c,p)-Faktorladung-Matrix mit c Faktorladungen für p Deskriptoren.
F ist die (n,p)-Residual-Matrix, die nicht erklärte Fehlerstreuung.

Dies entspricht der Anpassung einer c-dimensionalen Hyperfläche and die durch X definierte Punktewolke. Dabei entspricht die erste Hauptkomponente/Achse derjenigen Gerade im p-dimensionalen Raum, welche die Punktewolke am besten approximiert – im Sinne der kleinsten quadratischen Abweichung. Die zweite Hauptkomponente ist orthogonal zur ersten und approximiert die verbleibende Variation der Punktewolke am besten, usw. Die Faktorwerte T entsprechen den projizierten Koordinaten der Beobachtungen auf der c-dimensionalen Hyperfläche. Die Faktorladungen P entsprechen den Beiträgen der p Deskriptoren zu den c Hauptkomponenten und definieren geometrisch die Orientierung der c-dimensionalen Hyperfläche im Verhältnis zum ursprünglich p-dimensionalen Raum.

D.4 *Partial-Least-Squares-Regression* (PLS)

Mit Hilfe der *Partial-Least-Squares-Regression* (PLS) wird die Beziehung zwischen der Deskriptoren-Matrix X und der Aktivitäten-Matrix Y durch eine multivariates lineares statistisches Modell formuliert (Gleichung D:1). Analog zur PCA werden dazu sowohl aus X als auch aus Y jeweils c Komponenten extrahiert. Geometrisch entspricht dies einer Projektion der Punktewolken aus X und Y auf c-dimensionale Hyperflächen (Gleichungen D:3 - D:5).

$$X = TP + F \qquad (D:3)$$

$$Y = UQ + G \qquad (D:4)$$

$$Y = TQ + H \qquad (D:5)$$

Projektion der Deskriptoren-Matrix *X* und der Aktivitäten-Matrix *Y*

X, Y sind die (*n,p*)-Deskriptoren- bzw. (*n,m*)-Aktivitäten-Matrizen mit *n* Beobachtungen (Strukturen) und *p* Deskriptoren bzw. *m* Aktivitäten pro Struktur.
T, U sind die (*n,c*)-Faktorwert-Matrizen mit *c* Faktorwerten für *n* Beobachtungen (Strukturen).
P, Q sind die (*c,p*)- bzw. (*c,m*)-Faktorladung-Matrizen mit *c* Faktorladungen für *p* Deskriptoren bzw. *m* Aktivitäten.
F, G, H sind die entsprechenden Residual-Matrizen.

Diese *c* orthogonalen Komponenten (latente Variablen) werden dabei so gewählt, dass gleichzeitig beide Punktewolken *X* und *Y* gut approximiert werden (Gleichungen D:3 und D:4), sowie die Kovarianz zwischen den projizierten Punktewolken *T*(*X*) und *U*(*Y*) maximiert wird (Gleichung D:6). Auf Grund dieser inneren Relation gilt auch Gleichung D:5.

$$U = T + J \qquad (D:6)$$

Innere Relation zwischen *U* und *T*

Man beachte, dass die Koeffizienten der inneren Relation 1 sind. *J* ist die Residual-Matrix.

Im Rahmen des PLS-Algorithmus wird dazu eine Matrix *W* bestimmt, welche die Kovarianzstruktur zwischen *X* und *Y* repräsentiert. Über $T = XW$ wird anschließend *T* berechnet und erlaubt dadurch die Bestimmung von Q über Gleichung D:5. Über $B = WQ$ lassen sich die für Gleichung D:1 gesuchten PLS-Koeffizienten *B* für die *p* Deskriptoren bestimmen und damit das lineare Modell komplettieren.

Die optimale Anzahl an Komponenten wird sowohl bei der PCA als auch bei der PLS durch Kreuzvalidierung bestimmt. Dabei wird diejenige Anzahl an Komponenten für das finale Modell verwendet, welche den kreuzvalidierten Determinationskoeffizienten Q^2 maximiert. Durch PLS-Regression können sowohl im Rahmen der CoMFA- als auch der COMBINE-Methode – bei Vorliegen quantitativer Aktivitätsdaten – vorhersagefähige Modelle erstellt, welche ausgehend von der Struktur bzw. Komplexstruktur die Aktivität vorhersagen können.