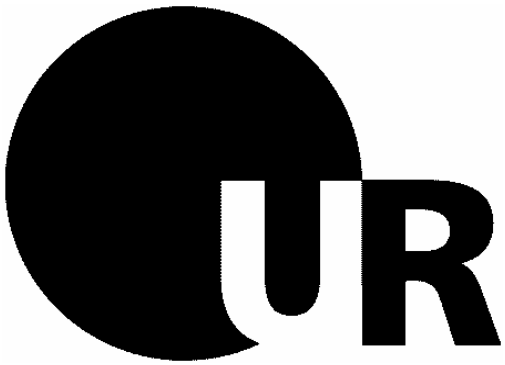
TRANSCENT– ein Enzymdesignprogramm zum Transfer aktiver Zentren unter Wahrung Katalyse-relevanter Rahmenbedingungen

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.) DER NATURWISSENSCHAFTLICHEN FAKULTÄT III – BIOLOGIE UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG



vorgelegt von
André Fischer aus Regensburg
im November 2007

Promotionsgesuch eingereicht am: 28.11.2007

Kolloquium fand statt am: 19.12.2007

Die Arbeit wurde angeleitet von: PD Dr. Rainer Merkl

Prüfungsausschuss:

Vorsitzender: Prof. Dr. Günter Hauska

Erstgutachter: PD Dr. Rainer Merkl

Zweitgutachter: PD Dr. Wolfram Gronwald

Drittprüfer: Prof. Dr. Reinhard Sterner

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AEE	L-Ala-D/L-Glu-Epimerase
Ala	Alanin
Arg	Arginin
Asn	Asparagin
Asp	Aspartat
Cys	Cystein
DEE	Dead Ends Elimination
Gln	Glutamin
Glu	Glutamat
Gly	Glycin
GMEC	Global Minimum Energy Conformation
His	Histidin
HisA	5'-ProFAR Isomerase
HisF	Imidazolglycerolphosphat Synthase
Ile	Isoleuzin
Leu	Leuzin
Lys	Lysin
Met	Methionin
MLE II	Mukonat laktonisierende Enzym II
MSA	Multiples Sequenzalignment
OMP	Orotidin 5'-Phosphat
OMPD	OMP Decarboxylase
OSBS	O-Succinylbenzoatsynthase
Phe	Phenylalanin
PRFAR	N´-[(5´-phosphoribulosyl)-formimino]-5-aminoimidazol-4-carboxamid-ribonukleotid
Pro	Prolin
PSA	Paarweises Sequenzalignment

RCdRP	reduziertes 1-(o-Carboxyphenylamino)-1-Desoxyribulose-5-Phosphat
SA	Simulated Annealing
Ser	Serin
Thr	Threonin
TIM	Triosephosphatisomerase
Trp	Tryptophan
TrpA	Alpha-Untereinheit der Tryptophansynthase
TrpC	Indolglycerinphosphat Synthase
TrpF	Phosphoribosylanthranilat Isomerase
Tyr	Tyrosin
Val	Valin

1 Kurzfassung

Nahezu alle Reaktionen des zellulären Stoffwechsels werden durch spezialisierte Proteine - die Enzyme - katalysiert, die meist enorme Effizienz aufweisen. Mit den heutigen Methoden der Biotechnologie ist es möglich, natürlich vorkommende Proteine zu verändern und völlig neuartige Proteinarchitekturen zu entwerfen. Allerdings ist es außerordentlich schwierig, Enzymfunktionen gezielt zu konstruieren. Ein weiterer, bisher nur in Ansätzen verstandener Aspekt von Enzymen ist deren Evolution. Für viele Enzyme ist unstrittig, dass sie von einem gemeinsamen Vorfahren abstammen. Allerdings kann die Entwicklung der verschiedenartigen Varianten höchstens in Einzelfällen im Detail verfolgt und erklärt werden.

Umwandlungsexperimente, bei denen das aktive Zentrum eines Enzyms auf das Proteingerüst eines anderen übertragen wird, können Einsichten in die evolutionäre Entwicklung spezifischer Enzymfunktionen liefern. Gleichzeitig erweitern solche Experimente und die hierbei gesammelten Erfahrungen das generelle Verständnis für Probleme des Enzymdesigns.

In dieser Arbeit wird das Programm TRANSCENT vorgestellt, das Modelle für Umwandlungsexperimente automatisch generiert. Grundlage der Modellierung sind die Raumstrukturen zweier Enzyme. Eine Struktur bildet die Vorlage und liefert die Informationen zur Enzymfunktion, die übertragen werden soll. Dazu gehören die Definition des aktiven Zentrums und die Lage des Liganden. Die zweite Struktur dient als Gerüst, auf dem diese enzymatische Funktion etabliert werden soll.

Beim Modellieren berücksichtigt das Programm vier verschiedene Rahmenbedingungen: 1) Hinreichende Stabilität des Proteinmodells. 2) Ligandenbindung im aktiven Zentrum. 3) Das Einstellen von Katalyse-relevanten pKa-Werten. 4) Die Ähnlichkeit des neuen aktiven Zentrums zur Vorlage.

Methoden um diese vier Rahmenbedingungen zu erfassen wurden als separate Module entwickelt und können in unterschiedlichen Kombinationen zur Optimierung der Umwandlung verwendet werden. Der Einfluss der Module auf die Qualität des Gesamtergebnisses wurde durch die Bewertung eines Testdatensatzes untersucht. Es wurde gezeigt, dass jedes einzelne Modul einen unabhängigen Beitrag zur optimalen Modellierung eines Transfers aktiver Zentren leistet.

Als Fallstudie wurde mit TRANSCENT die Übertragung aktiver Zentren zwischen verschiedenen $(\beta\alpha)_8$-Barrel-Proteinen aus der Histidin- und Tryptophan-Biosynthese modelliert. Die resultierenden Strukturmodelle wurden detailliert untersucht und auf Konsistenz überprüft.

2 Einleitung

Die DNA wird oft als Bauplan des Lebens bezeichnet. Dieser Vergleich ist sehr treffend, denn sie kodiert Gene, die Bauanleitungen für Proteine. Die Gene legen fest, in welcher Reihenfolge Aminosäuren miteinander zu einer Kette verknüpft werden. Von Ribosomen zusammengebaut, faltet sich diese Aminosäurekette selbständig zu einem Protein mit einer definierten Struktur und Funktion. Obwohl alle Proteine gleichartig aufgebaut sind und sich nur in ihrer Aminosäuresequenz voneinander unterscheiden, sind ihre Strukturen und Funktionen außerordentlich unterschiedlich. Viele Proteine katalysieren biochemische Reaktionen. Diese Reaktionen finden im aktiven Zentrum (Abbildung 1) statt, einer Vertiefung an der Proteinoberfläche. Solche Proteine heißen Enzyme. Es gibt eine enorme Vielzahl von Enzymen, die jeweils andere Reaktionen katalysieren. Damit bilden sie die Basis für den zellulären Stoffwechsel.

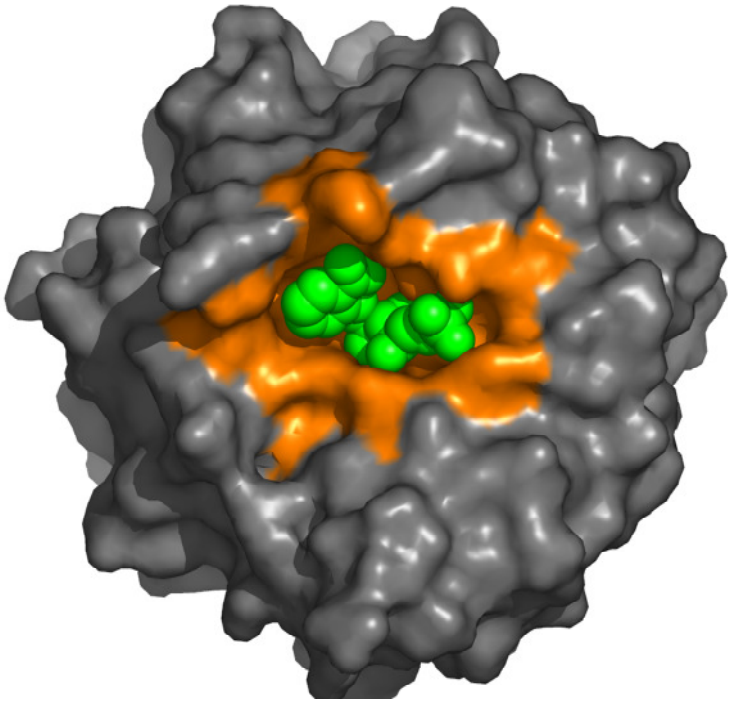


Abbildung 1: Enzym mit aktivem Zentrum und gebundenem Liganden
Das Bild zeigt die Oberflächenstruktur eines Enzyms (grau), das im aktiven Zentrum (orange) einen Liganden (grün) gebunden hat. Das Enzym ist die Phosphoribosylanthranilat Isomerase (TrpF), aus *Thermotoga maritima,* die einen Schritt der Tryptophan-Biosynthese katalysiert. Der Ligand ist reduziertes 1-(o-Carboxyphenylamino)-1-Desoxyribulose-5-Phosphat (rCdRP). Gut zu erkennen ist die Komplementarität der Bindetasche. Dieses „Schlüssel-Schloss"-Prinzip ist typisch für viele Enzyme und wurde schon 1894 von Emil Fischer beschrieben.

Es ist weitgehend unklar, wie im Laufe der Evolution Proteine mit so unterschiedlichen Funktionen entstanden sind und welche Änderungen an bestehenden Proteinen notwendig waren, um neue Funktionen zu generieren. Umwandlungsexperimente, bei

denen die katalytische Funktion von einem Enzym auf ein anderes übertragen wird, können Einblicke in diese Prozesse gewähren (Gerlt & Babbitt, 2001; Jürgens et al., 2000; Leopoldseder et al., 2004). Dies gilt, da die für eine Umwandlung notwendigen Austausche einen möglichen Pfad beschreiben, den auch die Evolution gewählt haben könnte, um eine neue Funktion hervorzubringen.

Die enorme funktionale Vielfalt der Proteine weckt auch Hoffnungen, dass es in der Zukunft gelingen könnte, neuartige Enzyme mit einer Funktion zu entwerfen, die in der Natur nicht vorkommen. Ganz im Sinne einer modernen Ingenieursdisziplin könnten sie zuerst am Computer modelliert, dann als Gene kodiert und schließlich von der natürlichen Synthesemaschienerie als Proteine synthetisiert werden. Erste Erfolge auf diesem Gebiet sind bereits publiziert (Bolon & Mayo, 2001; Dwyer et al., 2004). Künstliche Enzyme würden der Pharmazie und der Chemie ganz neue Möglichkeiten eröffnen. Auch hier können Umwandlungsexperimente wegweisend sein. Das Übertragen einer Enzymfunktion von einem Protein auf ein anderes kann als „*proof of principle*“ für das Etablieren einer neuen Funktion gesehen werden.

Eine effiziente Methode zum Transfer von Enzymfunktionen würde somit ein wertvolles Werkzeug für eine ganze Reihe von Anwendungsgebieten darstellen. Um diesem Ziel näher zu kommen, wurde im Rahmen dieser Arbeit ein Computerprogramm entwickelt, das automatisch die Änderungen ermittelt, die notwendig sind, um das aktive Zentrum eines Enzyms auf ein anderes Protein zu übertragen.

2.1 Die Umwandlung von Enzymen

Um die Evolution von verwandten Enzymen zu untersuchen, die unterschiedliche Funktionen entwickelt haben, kann man konkrete Paare von Enzymen auswählen und versuchen die Funktion des einen Enzyms durch die Funktion des anderen zu ersetzen (Abbildung 2). Dazu müssen die Enzyme analysiert und die Unterschiede zwischen ihnen herausgearbeitet werden. Schließlich kann eine Menge von Änderungen postuliert werden, die das eine Enzym so verändert, dass es die Funktion des anderen Enzyms übernimmt. Auf Sequenzebene können solchen Änderungen relativ einfach als Mutationen, sowie als Insertionen oder Deletionen modelliert werden. Die Herausforderung besteht darin, die Änderungen auf Strukturebene zu modellieren und die strukturellen Auswirkungen richtig abzuschätzen. Ein resultierendes Enzymmodell lässt sich überprüfen, indem die postulierten Änderungen im Labor nasschemisch umgesetzt werden. Lässt sich anschließend die Funktion des anderen Enzyms tatsächlich nachweisen, beschreiben die Änderungen einen erfolgreichen Umwandlungsprozess.

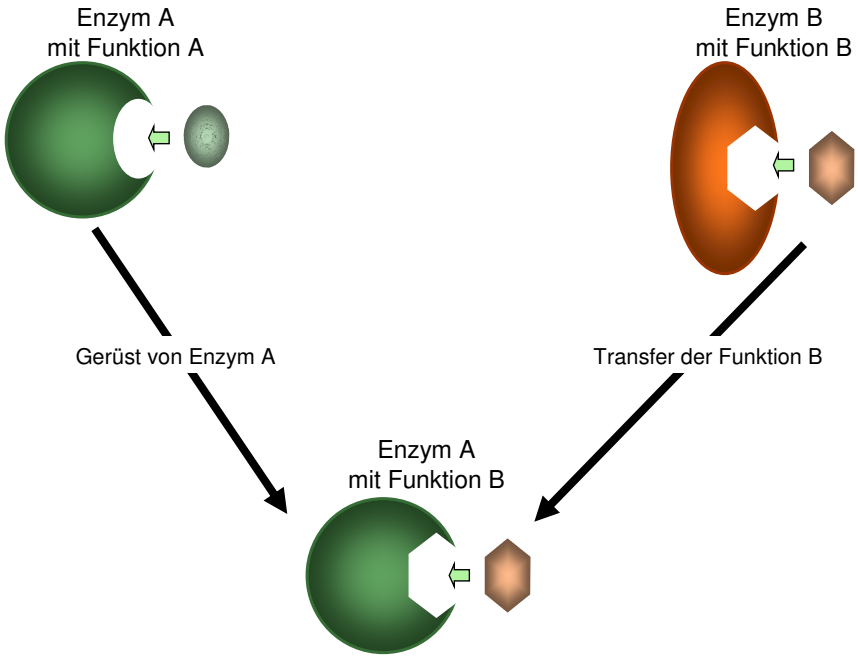


Abbildung 2: Umwandlung von Enzymen
Das Schema verdeutlicht einen Umwandlungsprozess für zwei Enzyme. Enzym A kann in sein aktives Zentrum nur das Substrat für die Reaktion A aufnehmen, Enzym B nur das Substrat für die Reaktion B. Diese Spezifität ist durch die Form der aktiven Zentren angedeutet (Schlüssel-Schloss-Prinzip). Für die Umwandlung liefert das Enzym A das Proteingerüst, auf das die Funktion B des Enzyms B transferiert wird. Dabei geht die ursprüngliche Funktion von Enzym A verloren. Transferieren bedeutet hierbei, das aktive Zentrum von Enzym A durch Mutationen so anzupassen, dass es dem aktiven Zentrum von Enzym B entspricht.

Die Menge der Änderungen sollte möglichst klein sein. Wenn diese Menge alle Änderungen enthält, mit der die Sequenz des einen Enzyms in die Sequenz des anderen überführt wird, dann ist die Fragestellung nach der Evolution *ad absurdum* geführt. Denn es sind vor allem diejenigen Änderungen interessant, die zwingend für die neue Funktion erforderlich sind. Eine Vorhersage sollte sich daher auf diese Austausche beschränken.

Um Umwandlungsexperimente zwischen Enzymen erfolgreich durchführen zu können, ist es von Vorteil, wenn eine Reihe von Voraussetzungen erfüllt ist. Zunächst muss die 3D-Struktur der beiden Enzyme aufgeklärt sein. Alternativ sollten zumindest zuverlässige Computermodelle (Ginalski, 2006) der Strukturen zur Verfügung stehen. Nur so lässt sich der Einfluss der Änderungen auf struktureller Ebene analysieren. Weiterhin sollten die Enzyme in Bezug auf ihre Sequenzähnlichkeit nicht zu unterschiedlich sein. Sonst wird es schwierig, das umgewandelte Enzym als Modell für natürliche Proteinevolution zu diskutieren (Sander & Schneider, 1991). Zusätzlich sollten die betrachteten Funktionen biochemisch gut verstanden sein. Dies impliziert, dass die für die Katalyse essentiellen Reste identifiziert sind und der Katalysemechanis-

mus aufgeklärt ist (siehe Beispiel in Abbildung 3). Nur so lassen sich die Änderungen auch in Bezug auf die Funktion interpretieren. Schließlich sollten die Proteine experimentell gut handhabbar sein. Damit ist gemeint, dass sich die Proteine gut aufreinigen lassen und dass ein biochemischer Test für ihre Funktion existiert.

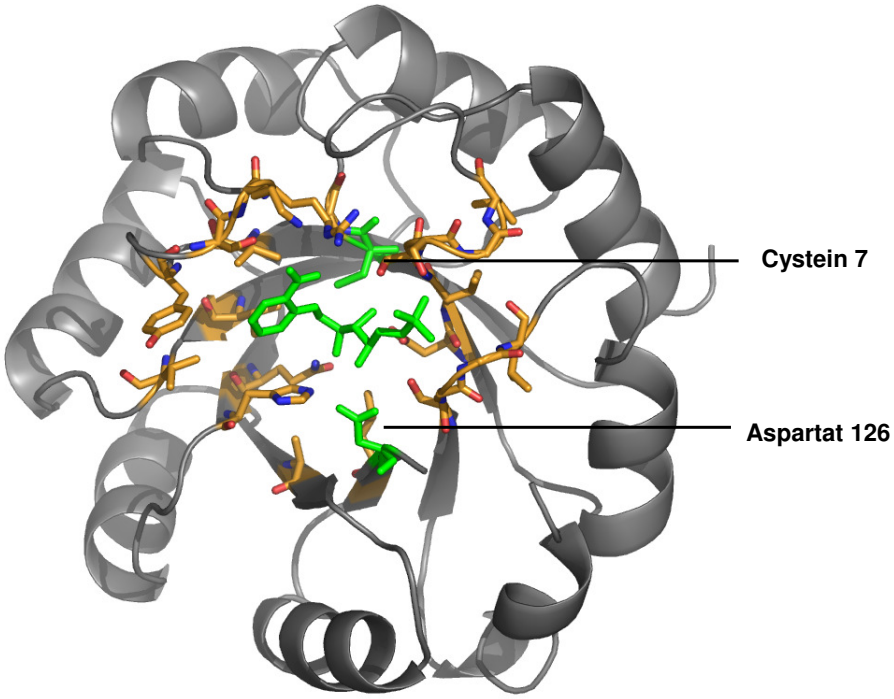


Abbildung 3: Katalytisch essentielle Reste eines aktiven Zentrums
Das aktive Zentrum des abgebildeten Enzyms TrpF aus *T. maritima* besteht aus ca. 30 Resten (orange). Die meisten Reste sind an der Bindung des Liganden rCdRP (grün, Mitte) beteiligt. Für die Katalyse verantwortlich sind nur zwei Reste: Cystein 7 (grün, oben) und Aspartat 126 (grün, unten). Der Katalysemechanismus ist eine Amadori-Umlagerung, wobei das Aspartat als allgemeine Base und das Cystein als allgemeine Säure dient (Henn-Sax et al., 2002).

Zur Funktionsumwandlung von Enzymen stehen zwei, in der Vorgehensweise völlig unterschiedliche Strategien zur Verfügung (Glasner et al., 2007; Johannes & Zhao, 2006; Leisola & Turunen, 2007): Gelenkte Evolution und rationales Design.

Gelenkte Evolution beruht auf den Prinzipien der Zufallsmutagenese und der Selektion. Bei der Zufallsmutagenese wird durch fehlerhaftes Kopieren (*error prone PCR*) eine große Menge unterschiedlicher Genvarianten des Proteins erzeugt, dessen Funktion umgewandelt werden soll. Diese Menge der Varianten wird Genbank genannt.

Durch einen nachfolgenden Screening- oder Selektionsprozess werden dann Varianten mit den gewünschten Eigenschaften aus der Genbank isoliert. Ein solcher Selektionsprozess ist zum Beispiel der Komplementationstest. Dazu werden die verschiedenen Varianten aus der Genbank in Bakterienzellen geschleust, denen die gesuchte katalytische Funktion fehlt. Anschließend werden die Zellen Bedingungen ausgesetzt, die sie

nur überleben können, wenn sie mit Hilfe eines anderen Proteins die fehlende Funktion ersetzen (d.h. komplementieren) können. Wenn Zellen überleben, tragen diese folglich ein eingeschleustes Genprodukt mit der gesuchten Funktion in sich.

Im Gegensatz zu diesem zufallsbasierten Ansatz werden beim rationalen Design Informationen über das Protein und die zu etablierende Funktion zusammengetragen, um daraus die notwendigen Änderungen für die Umwandlung abzuleiten. Solche Informationen sind zum Beispiel Proteinstrukturen, Sequenzalignments oder das Wissen über den Katalysemechanismus. Das rationale Design wird entweder manuell oder computergestützt durchgeführt. Die vorgeschlagenen Änderungen werden dann mit Hilfe von positionsspezifischer Mutagenese in das Protein eingeführt. Beispiele für rationales Design sind Konsensusdesigns auf der Basis von Multiplen Sequenzalignments (Russ et al., 2005) oder computerbasiertes Proteindesign auf Basis von Proteinstrukturen (Bolon & Mayo, 2001) (Dwyer et al., 2004).

2.2 $(\beta\alpha)_8$-Barrel-Enzyme

Die $(\beta\alpha)_8$-Barrel Enzyme bilden eine besonders große Klasse von Proteinen (Nagano et al., 2002; Sterner & Höcker, 2005; Wierenga, 2001). Diese Enzyme sind einerseits strukturell sehr ähnlich, andererseits in Bezug auf ihre Enzymfunktion enorm vielseitig. Da die strukturelle Ähnlichkeit eine evolutionäre Verwandtschaft vermuten lässt, sind $(\beta\alpha)_8$-Barrel-Proteine ein interessantes Modellsystem, um Proteinevolution in Bezug auf die Entwicklung neuer Funktionen zu untersuchen. Die enorme Funktionsvielfalt lässt vermuten, dass gerade dieser Faltungstyp geeignet ist, neue Enzymfunktionen zu entwickeln.

2.2.1 Die Faltung und Evolution von $(\beta\alpha)_8$-Barrel-Proteinen

$(\beta\alpha)_8$-Barrel sind globuläre Proteine, die jeweils aus etwa 250 Aminosäuren bestehen. Der $(\beta\alpha)_8$-Barrel-Faltungstyp ist sehr häufig und vielseitig. In der SCOP Datenbank (Murzin et al., 1995) umfasst er 32 Superfamilien, 91 Familien und 232 verschiedene $(\beta\alpha)_8$-Barrel Domänen.

Charakteristisch für die Faltung ist ein sich achtmal wiederholendes Motiv aus einem β-Strang und einer α-Helix. Im Zentrum lagern sich die β-Stränge zu einem ringförmigen parallelen β-Faltblatt zusammen. Dieser gleichmäßigen, fassförmigen Struktur verdankt die $(\beta\alpha)_8$-Barrel-Faltung ihren Namen (Abbildung 4). Die Topologie wird auch TIM-Barrel genannt, da sie zum ersten Mal bei der Triosephosphat Isomerase gefunden wurde (Banner et al., 1975).

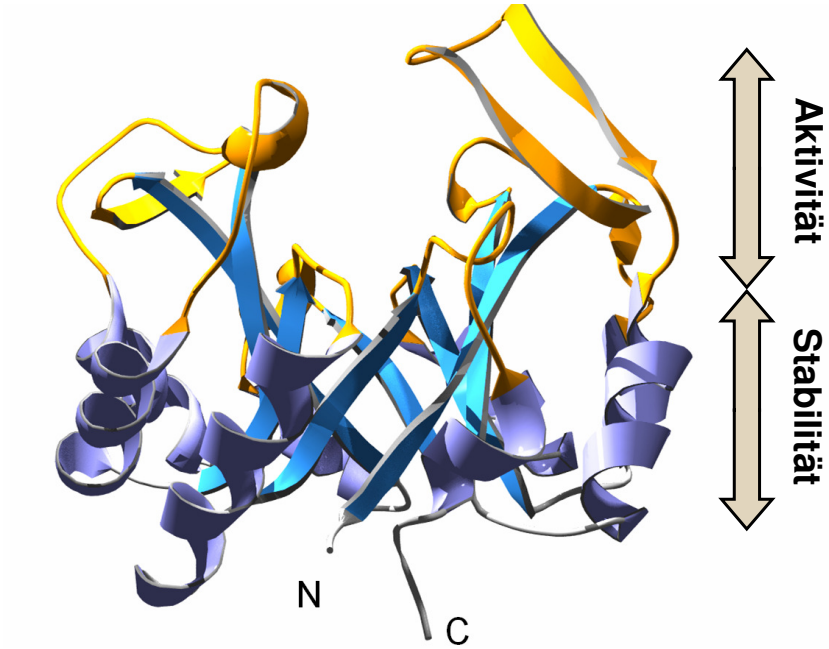


Abbildung 4: Die $(\beta\alpha)_8$-Barrel Topologie

Die Abbildung zeigt die Imidazolglycerolphosphat Synthase (HisF), ein typisches $(\beta\alpha)_8$-Barrel-Protein aus der Histidin-Biosynthese. Um das für die $(\beta\alpha)_8$-Barrel charakteristische zentrale β-Faltblatt (hellblau) herum ordnen sich die 8 α-Helices an (violett). Dieser Bereich wird als für die Stabilität besonders wichtig erachtet. Das aktive Zentrum befindet sich in den Schleifenregionen oberhalb des zentralen β-Faltblatts (gelb). In dieser Region liegt bei allen bisher bekannten $(\beta\alpha)_8$-Barrel-Enzymen das aktive Zentrum.

Die generell hohe strukturelle Übereinstimmung und eine zwischen einzelnen Vertretern noch deutlich vorhandene Sequenzähnlichkeit legen den Schluss eines gemeinsamen evolutionären Ursprungs der $(\beta\alpha)_8$-Barrel-Enzyme nahe (Henn-Sax et al., 2001; Lang et al., 2000). Dieser Umstand lässt vermuten, dass die bekannten $(\beta\alpha)_8$-Barrel-Enzyme durch divergente Evolution aus einem gemeinsamen Vorläufer entstanden sind. Man nimmt außerdem an, dass die $(\beta\alpha)_8$-Barrel-Faltung eine sehr alte und ursprüngliche Faltung ist (Caetano-Anolles et al., 2007). Diese Annahme wird durch zwei Beobachtungen gestützt: Zum einen sind $(\beta\alpha)_8$-Barrel in essentiellen Biosynthesewegen, wie in der Aminosäurebiosynthese zu finden. Zum anderen decken sie fast alle wichtigen Enzymklassen durch eine große Menge an Vertretern ab. Das Entstehen derart unterschiedlicher Enzymfunktionen setzt einen gewissen Zeitraum für deren Evolution voraus.

$(\beta\alpha)_8$-Barrel-Enzyme sind in 5 der 6 EC-Klassen vertreten (Nagano et al., 2002; Pujadas et al., 1996). Eine Ausnahme bilden lediglich die Ligasen. Diese außerordentliche Vielfalt ist erklärbar durch das hohe Alter, das für diesen Faltungstyp angenommen wird (Caetano-Anolles et al., 2007). Die Vielfalt spricht aber auch für die

universelle Eignung des Faltungstyps für enzymatische Funktionen (Wierenga, 2001). Dies wird besonders an der uniformen Topologie der $(\beta\alpha)_8$-Barrel-Enzyme deutlich, denn bei allen Vertretern befindet sich das aktive Zentrum über dem Ende des zentralen β-Faltblatts, das durch die C-Termini der β-Stränge gebildet wird. Diese Region wird Katalysepol (Sterner & Höcker, 2005) genannt und besteht aus den C-terminalen Enden der β-Stränge und den Schleifen, welche die β-Stränge mit den nachfolgenden α-Helices verbinden. Die große funktionale Vielfalt ist mit hoher Variabilität in diesen Schleifenregionen verbunden. Die Schleifen verschiedener $(\beta\alpha)_8$-Barrel-Enzyme variieren dabei in ihrer Länge und Aminosäuresequenz, was dem Katalysepol Plastizität verleiht. Die Region am anderen Ende des zentralen β-Faltblatts wird Stabilitätspol genannt und ist in vielen $(\beta\alpha)_8$-Barreln für Stabilität verantwortlich. Diese topologische Unterteilung macht die Funktionsvielfalt sehr plausibel, denn durch die räumliche Trennung von Stabilität und Funktion sind diese beiden Aspekte voneinander weitgehend entkoppelt.

Der topologische Aufbau der $(\beta\alpha)_8$-Barrel-Enzyme lässt vermuten, dass sich neue Funktionen durch wenige Änderungen aus vorhandenen Funktionen entwickelt haben. Daher stellt dieser Faltungstyp ein interessantes Modellsystem für die Proteinevolution dar. An einzelnen Beispielen sollte sich dann besonders einfach nachvollziehen lassen, durch welche Anpassungen sich neue Funktionen entwickelt haben.

Treffen diese Annahmen zu, so hätte diese Erkenntnis große Bedeutung für die moderne Biotechnologie. Die $(\beta\alpha)_8$-Barrel-Faltung würde sich als universelles Grundgerüst für die Entwicklung künstliche Enzyme empfehlen. Es sollte dann auch technisch einfacher sein, neue Funktionen auf $(\beta\alpha)_8$-Barrel-Proteinen als auf Proteinen mit anderen Faltungstypen zu etablieren. Mit Umwandlungsexperimenten lässt sich diese Hypothese genauer untersuchen.

2.2.2 Die Superfamilie der Ribulosephosphat-bindenden Barrel

Aufgrund der großen Menge an verschiedenen $(\beta\alpha)_8$-Barrel-Enzymen sind sehr viele Paare für Umwandlungsexperimente denkbar. Es ist sinnvoll, mit einfachen Paaren zu beginnen, die gute Voraussetzungen für einen Umwandlungserfolg haben. Interessante Kandidaten für Umwandlungsexperimente finden sich in der Superfamilie der Ribulosephosphat-bindenden Barrel-Enzyme. Das sind $(\beta\alpha)_8$-Barrel-Enzyme, die jeweils ein Substrat mit einem Ribulosephosphat-Rest umsetzen. Die Superfamilie wiederum besteht aus fünf Familien, von denen sich zwei Familien in besonderer Weise eignen. Dies sind die $(\beta\alpha)_8$-Barrel-Enzyme der Histidin-Biosynthese (Alifano et al., 1996; Charlebois et al., 1997) und der Trypthophan-Biosynthese (Yanofsky, 2001; Yanofsky, 2003). Die Vertreter aus der Histidin-Biosynthese sind die 5'-ProFAR

Isomerase (HisA) und die Imidazolglycerolphosphat Synthase (HisF). Die Vertreter aus der Trypthophan-Biosynthese sind die α-Untereinheit der Tryptophansynthase (TrpA), die Indolglycerinphosphat Synthase (TrpC) und die Phosphoribosylanthranilat Isomerase (TrpF).

Alle Enzyme weisen eine gemeinsames Phosphatbindemotiv an den Enden des siebten und achten β-Strangs auf. Über diese Phosphatbindestelle sind in diesen Enzymen die Substrate gleichartig ausgerichtet. HisA und HisF sowie TrpF, TrpC und TrpA katalysieren aufeinander folgende Schritte in ihren Biosynthesewegen. Daher sind die Substrate zum Teil auch über den Ribulosephosphat-Rest hinaus strukturell ähnlich (Sterner & Höcker, 2005).

Histidin- und Trypthophan-Biosynthesewege sind sehr gut untersucht und verstanden. So ist für jedes der genannten Enzyme die 3D-Struktur in höherer Auflösung bekannt und ein Funktionstests verfügbar.

2.2.3 Der Status von $(\beta\alpha)_8$-Barrel Umwandlungen

Mit $(\beta\alpha)_8$-Barrel-Enzymen wurde schon eine Reihe von Umwandlungen versucht. Die im Folgenden aufgeführten Experimente (siehe Tabelle 1) betreffen HisA, HisF, TrpA, TrpC, TrpF und die TIM (Altamirano et al., 2000; Altamirano et al., 2002; Darimont et al., 1998; Dwyer et al., 2004; Henn-Sax et al., 2002; Jürgens et al., 2000; Lambeck, 2004; Leopoldseder et al., 2004; Riede, 2006), sowie $(\beta\alpha)_8$-Barrel-Enzyme aus der Enolase-Superfamilie (Schmidt et al., 2003):

1. In zwei erfolgreichen Umwandlungen (Jürgens et al., 2000; Leopoldseder et al., 2004) konnte die katalytische Aktivität von TrpF auf HisA und HisF etabliert werden. Dazu war in beiden Fällen an analogen Positionen des aktiven Zentrums nur der Austausch eines Aspartates durch ein Valin, bzw. eine andere nicht negativ geladenen Aminosäure notwendig. Diese Mutation wurde durch gerichtete Evolution gefunden. Die wildtypischen Aktivitäten waren durch den Austausch in beiden Fällen stark reduziert. Auch wenn in beiden Fällen die neuen Enzyme fast nur noch TrpF-Aktivität aufweisen, unterscheiden sich die aktiven Zentren der Mutanten deutlich von TrpF. Die katalytisch essentiellen Reste von TrpF, ein Aspartat und ein Cystein (vgl. Abbildung 3), wurden weder nach HisA und HisF übertragen, noch waren sie dort bereits vorhanden. Welche Bedeutung der Valin-Austausch in Bezug auf gemeinsame Evolution von HisA, HisF und TrpF hat, ist daher schwierig zu klären.
2. Zumindest für HisA kann wildtypische Aktivität auch gemeinsam mit TrpF-Aktivität in einem $(\beta\alpha)_8$-Barrel etabliert werden. In (Kuper et al., 2005; Wright et al., 2007) wurde die Struktur des bifunktionalen Enzyms PriA aufgeklärt, das

homolog zu HisA ist und beide Reaktionen mit hoher Effizienz katalysiert (Barona-Gomez & Hodgson, 2003). Erstaunlicherweise ist das Aspartat, das in HisA und HisF für die TrpF-Aktivität zu Valin mutiert werden musste, in PriA an äquivalenter Position vorhanden. Ob der Valin-Austausch einen wichtigen Schritt in der Evolution zwischen HisA und TrpF darstellt, lässt sich daher auch unter Einbeziehung von PriA nicht einfach erklären.

3. In zwei Fällen wurden versucht, die Aktivität von TrpF auf TrpC zu übertragen (Darimont et al., 1998), indem die katalytisch essentiellen Reste von TrpF an äquivalenten Positionen in TrpC eingeführt wurden. Die äquivalenten Positionen wurden dabei durch ein strukturbasierten Sequenzalignment bestimmt. Dieser Ansatz war weder für TrpC von *Escherichia coli* noch für TrpC aus *Sulfolobus solfataricus* (Darimont et al., 1998) erfolgreich. In beiden Fällen zeigten die Mutanten weder TrpF- noch die ursprüngliche TrpC-Aktivität.

4. Die Umwandlung von TrpC aus *E. coli* zu TrpF wurde in (Altamirano et al., 2000) wiederholt. In dieser Arbeit wurden jedoch ganze Sequenzfragmente in TrpC deletiert und dafür Fragmente aus TrpF eingeführt. Dieses Konzept wurde mit gerichteter Evolution kombiniert. Die in der Arbeit publizierte TrpF-Aktivität des finalen Konstruktes wurde später zurückgezogen (Altamirano et al., 2002). Vermutlich lag eine Kontamination mit wildtypischem TrpF vor.

5. Die Umwandlung von TrpC aus *S. solfataricus* zu TrpF wurde in der Diplomarbeit von Philipp Riede (Riede, 2006) wiederholt und um Mutationen zur pKa-Optimierung der katalytischen Reste erweitert. Dazu wurde mit dem Programm PROPKA (Li et al., 2005) ein Netz von Resten identifiziert, das die pKa-Werte der katalytisch essentiellen Reste von TrpF beeinflusst. Zusammen mit den essentiellen Resten wurden die wichtigsten Reste dieses Netzes auf TrpC übertragen. Die resultierende Mutante konnte nicht löslich aufgereinigt und daher auch nicht vermessen werden. Zusätzlich wurde mit gerichteter Evolution versucht, TrpF-Aktivität auf TrpC zu etablieren. Dabei konnten keine aktiven Varianten gefunden werden.

6. In (Lang et al., 2000) wurde gezeigt, dass bereits wildtypisches HisF aus *Thermotoga maritima* schwache HisA-Aktivität aufweist. Es ist anzunehmen, dass diese promiskuitive Aktivität (Khersonsky et al., 2006) durch geeignete Mutationen gesteigert werden kann.

7. In der Diplomarbeit von Iris Lambeck (Lambeck, 2004) wurde versucht, die TIM-Aktivität auf TrpA von *T. maritima* zu etablieren. Dazu wurden die katalytischen Reste von TIM über ein strukturbasiertes Sequenzalignment auf TrpA übertragen. Die so erzeugten Mutanten zeigten keine TIM-Aktivität. Parallel dazu wurde mit

gerichteter Evolution versucht, TIM-Aktivität auf TrpA zu etablieren. Auch diese Experimente blieben ohne Erfolg.

8. Die Enolase-Superfamilie besteht aus $(\beta\alpha)_8$-Barrel-Enzymen. Vertreter sind die L-Ala-D/L-Glu-Epimerase (AEE), das Mukonat laktonisierende Enzym II (MLE II) und die O-Succinylbenzoatsynthase (OSBS). Durch rationales Design wurde in AEE von *E. coli* ein Austausch (Aspartat zu Glycin) identifiziert, der zu einer OSBS-Nebenaktivität führt. Durch gerichtete Evolution wurde auch für MLE II von *Pseudomonas* sp. P51 ein ähnlicher Austausch gefunden (Glutamat zu Glycin an äquivalenter Position), der ebenfalls zu einer OSBS-Nebenaktivität führt (Schmidt et al., 2003).

9. Das folgende Umwandlungsexperiment unterscheidet sich aufgrund des Strukturgerüsts von den anderen Experimenten dieser Aufzählung. In (Allert et al., 2007; Dwyer et al., 2004) konnte die TIM-Aktivität auf das Ribosebindeprotein übertragen werden, welches aber, im Gegensatz zur TIM, kein $(\beta\alpha)_8$-Barrel-Protein ist. Die Umwandlung wurde durch ein Computerprogramm (DEZYMER) modelliert. Grundlage für die Modellierung bildeten die 3D-Strukturen der beiden Proteine. Das Programm übertrug zuerst die katalytischen Reste der TIM in die Bindetasche des Ribosebindeproteins. Anschließend wurde die Bindetasche durch weitere Austausche für die optimale Substratbindung angepasst. Im Labor konnte dann tatsächlich eine basale TIM-Aktivität nachgewiesen werden. Diese wurde anschließend durch gerichtete Evolution noch gesteigert.

Wie in Kapitel 2.2.2 bereits ausgeführt, sind für Umwandlungen von $(\beta\alpha)_8$-Barrel-Proteinen eine Reihe günstiger Voraussetzungen gegeben. Auch die Existenz von wildtypischen $(\beta\alpha)_8$-Barrel mit zwei Funktionen, wie PriA (HisA, TrpF) und HisF (HisF, HisA) legen Erfolgsaussichten für Umwandlungen nahe. Die Misserfolge in den diskutierten Beispielen machen aber deutlich, dass solche Experimente keineswegs von trivialer Natur sind.

Bei den hier betrachteten Umwandlungen von $(\beta\alpha)_8$-Barrel-Proteinen war gerichtete Evolution die erfolgreichere Strategie. In den Fällen von (Jürgens et al., 2000; Leopoldseder et al., 2004) (Schmidt et al., 2003) war für die Umwandlung nur ein Austausch notwendig, der mit gerichteter Evolution auch gefunden wurde. In allen anderen Fällen führte die Verwendung dieser Strategie nicht zum Ziel. Es muss daher davon ausgegangen werden, dass für diese Fälle die benötigten Austausche in den verwendeten Genbanken bei der Selektion nicht gefunden wurden oder nicht vorhanden waren. Die Wahrscheinlichkeit in einer Genbank eine neue Aktivität zu finden, sinkt mit der Anzahl der benötigten Austausche (Johannes & Zhao, 2006). Da normalerweise nicht *a priori* klar ist, wie viele Austausche für eine Umwandlung mindestens benötigt werden, sind die Erfolgsaussichten solcher Experimente nur schwer abzuschätzen.

Tabelle 1: Bisher durchgeführte Umwandlungsexperimente für $(\beta\alpha)_8$-Barrel-Proteine

Wildtyp Aktivitäten sind mit *wt*, erfolgreiche Umwandlungen mit √ und erfolglose Umwandlungen mit X gekennzeichnet. Die Experimente sind beschrieben in: 1 - (Lang et al., 2000), 2 - (Barona-Gomez & Hodgson, 2003) , 3 - (Jürgens et al., 2000), 4 - (Leopoldseder et al., 2004), 5, 6 - (Darimont et al., 1998), 7 - (Riede, 2006), 7 - (Altamirano et al., 2000; Altamirano et al., 2002), 9 - (Lambeck, 2004).

	Etabliert auf					
	HisA	HisF	TrpA	TrpC	TrpF	TIM
HisA-Aktivität	-	*wt* [1]				
HisF-Aktivität		-				
TrpA-Aktivität			-			
TrpC-Aktivität				-		
TrpF-Aktivität	*wt* [2], √[3]	√[4]		X[5,6,7,8]	-	
TIM-Aktivität			X[9]			-

Rationales Design auf Basis von strukturbasierten Sequenzalignments brachte in einem Experiment den Erfolg (Schmidt et al., 2003). In diesem Fall wurde die Substratspezifität eines Enzyms geändert. In den anderen Fällen sollte der Katalysemechanismus übertragen werden. Dabei war das alleinige Übertragen der katalytisch essentiellen Reste in keinem der Fälle ausreichend. Da diese Methode in der Menge der Austausche nicht beschränkt ist, wurden im Fall von (Riede, 2006) zusätzliche Mutationen eingeführt. Das resultierende Protein ließ sich allerdings nicht mehr aufreinigen. Dieses Resultat macht die Grenzen des Enzymdesigns deutlich: Führen neu eingeführte Reste zur Destabilisierung des Proteins, so lässt sich der Effekt der Mutation auf die Funktion nicht mehr untersuchen. Wenn also bei einer Umwandlung die Umgebung, in der die neuen Reste eingebettet werden, nicht berücksichtigt wird, so ist der Erfolg von vorne herein gefährdet. Durch Austausche eingeführte konformationelle Spannungen, das Auffinden weiterer kompensierender Austausche und das Abwägen zwischen verschiedenen Möglichkeiten bringt manuelles rationales Proteindesign schnell an seine Grenzen. Für die Umwandlung von (Dwyer et al., 2004) wurden diese Aufgaben von einem Computerprogramm übernommen, das die optimale Einbettung aller Austausche auf struktureller Basis berechnet hat. Auch wenn durch die Berechnungen des Computerprogramms eine Basisaktivität gefunden wurde, war die Modellierungsgenauigkeit nicht ausreichend, um diese Aktivität weiter zu optimieren. Dies konnte erst durch gerichtete Evolution erreicht werden.

Die Kopplung von rationalem Design und gerichteter Evolution scheint also für Umwandlungen eine ideale Kombination zu sein: Mit rationalem Design werden wichtige Reste positioniert und durch gerichtete Evolution werden anschließend weniger offensichtliche Änderungen gefunden, die für das gewünschte Ergebnis ebenfalls notwendig sind. Werden für das rationale Design Computerprogramme verwendet, können Umwandlungsmodelle auch in atomistischen Details optimiert werden, was die Erfolgsausichten weiter steigert.

2.3 Computermethoden

Enzyme sind Makromoleküle mit einem außerordentlich komplexen Wechselwirkungsgefüge zwischen den einzelnen Atomen. Um mit rationalem Design ein aktives Zentrum erfolgreich von einem Enzym auf ein anderes zu transferieren, ist es daher besonders viel versprechend, Computerprogramme einzusetzen, die den Transfer modellieren. Um die Auswirkungen des Transfers in den atomistische Details richtig zu beschreiben, wird ein Rahmenkonzept notwendig, welches erlaubt, die dreidimensionale Struktur von Enzymen zu modellieren und zu manipulieren.

Bei der Modellierung des Transfers sollten eine Reihe von Rahmenbedingungen berücksichtigt werden, die miteinander verknüpft sind und sich gegenseitig beeinflussen. Zu diesen Bedingungen zählen: Stabilität, Ligandenbindung, Ähnlichkeit des transferierten aktiven Zentrums zur Vorlage und die katalytische Reaktivität. Diese Rahmenbedingungen lassen sich wie folgt motivieren:

1. Damit ein Protein in Lösung eine definierte, gefaltete Struktur hat, muss es stabil sein. Stabilität bedeutet, dass die gefaltete Form energetisch günstiger ist, als die ungefaltete. Wenn für ein Protein ein neues aktives Zentrum modelliert werden soll, wird vorausgesetzt, dass sich die Proteinstruktur dabei nicht entfaltet. Nur unter dieser Annahme können gezielte Änderungen modelliert werden. Ein umfassender Überblick über die strukturellen Grundlagen der Proteinstabilität findet sich in (Jaenicke, 2000).
2. Die Ligandenbindung ist eine weitere wichtige Bedingung für einen erfolgreichen Transfer eines aktiven Zentrums, da Katalyse Bindung voraussetzt. Schon 1913 erkannte Paul Ehrlich: „*Corpora non agunt nisi fixata*“ – „keine Wirkung ohne Bindung“. Eine weitere, mit der Bindung verknüpfte Bedingung für die Katalyse ist das Schlüssel-Schloss-Prinzip, das Emil Fischer bereits 1894 beschrieb. Ligand und aktives Zentrum müssen zueinander komplementär sein, damit der Ligand in der richtigen Orientierung bindet (vgl. Abbildung 1). Nur wenn der Ligand und die katalytischen Reste richtig zueinander angeordnet sind, kann Katalyse stattfinden.

Gohlke und Klebe haben in einem Übersichtsartikel Strategien beschrieben, wie sich Ligandenbindung in Proteinen modellieren lässt (Gohlke & Klebe, 2002).

3. Ein übertragenes aktives Zentrum ist besonders plausibel, wenn es der Vorlage in Bezug auf alle relevanten Merkmale ähnlich ist (Gherardini et al., 2007). Relevante Merkmale sind zum Beispiel die Reste, die am Katalysemechanismus oder an der Ligandenbindung beteiligt sind. Da die Strukturen der Proteine, zwischen denen das aktive Zentrum transferiert wird, in der Regel nicht identisch sind, lassen sich die relevanten Reste nicht einfach eins-zu-eins übertragen. An welchen alternativen Stellen die Reste ebenfalls geeignet positionierbar sind, hängt vom dem Spielraum ab, der den Resten für ihre Funktion zur Verfügung steht. Damit ein neues aktives Zentrum auf Ähnlichkeit zu seiner Vorlage optimiert werden kann, müssen die relevanten Merkmale der Vorlage bestimmt werden. Zusätzlich müssen auch die Spielräume bestimmt werden, die den Merkmalen bei der Übertragung zur Verfügung stehen.
4. Der Transfer von Protonen zwischen Protein und Substrat ist Teil vieler enzymatischer Reaktionen. Damit ein katalytischer Rest ein Proton annehmen oder abgeben kann, muss er einen geeigneten pKa-Wert aufweisen, der in Enzymen oft erst durch den elektrostatischen Einfluss benachbarter Aminosäuren erreicht wird (Warshel et al., 2006). Damit ein aktives Zentrum erfolgreich übertragen werden kann, sollten im Idealfall auch die pKa-Werte essentieller Reste modelliert werden können.

Im Rahmen dieser Arbeit wurde das Computerprogramm TRANSCENT entwickelt, dass die Übertragung aktiver Zentren modelliert und dabei alle oben beschriebenen Bedingungen berücksichtigt. In den folgenden Kapiteln wird der Stand der Technik von Methoden zusammengefasst, die einen Bezug zur vorliegenden Arbeit haben. Die Erläuterungen dienen dazu, die Auswahl der verwendeten Berechnungsverfahren plausibel zu machen und den Aufbau von TRANSCENT zu erläutern.

2.3.1 Proteinstabilität durch computerbasiertes Proteindesign

Stabilität ist eine grundlegende Voraussetzung für Proteine. Nur wenn Proteine stabil sind, behalten sie ihre Struktur und die daran geknüpfte Funktion. Stabilität ist eine relativ gut verstandene Eigenschaft (Eijsink et al., 2004). Daher gibt es nicht nur Programme, die die Stabilität von Proteinen beurteilen (Guerois et al., 2002), sondern auch welche zur Optimierung der Proteinstabilität. Ein Ansatz für die Stabilitätsoptimierung ist die Lösung des inversen Faltungsproblem (Bowie et al., 1991). Beim inversen Faltungsproblem wird für ein gegebenes Proteinrückgrat eine optimale Aminosäuresequenz gesucht, welche die Struktur maximal stabilisiert. Den

Gegensatz hierzu bildet das Faltungsproblem (Dill et al., 2007), bei dem für eine gegebene Sequenz die stabilste Struktur gesucht wird. In der Literatur hat sich für das inverse Faltungsproblem der Begriff „Proteindesign“ etabliert, auch wenn dieser sich eigentlich nicht exklusiv auf Computermethoden beschränkt. Auch im folgenden Text bezeichnet von nun an der Begriff Proteindesign das inverse Faltungsproblem.

Proteindesign ist sehr komplex (Pierce & Winfree, 2002). Der Suchraum besteht aus der Menge aller möglichen Sequenzen für eine Struktur. Bereits bei einem Proteingerüst mit 100 Aminosäurepositionen ergeben sich 20^{100}, also etwa 10^{130} mögliche Sequenzen (zum Vergleich, die Anzahl der Atome im Weltall beträgt ca. 10^{68}). Da die meisten Seitenketten der einzelnen Aminosäuren zusätzliche Freiheitsgrade besitzen, wird der Suchraum noch größer.

Typische Proteindesignprogramme (Dahiyat & Mayo, 1997) bestehen aus drei Komponenten: einer Modellierungseinheit, um Proteinmodelle zu generieren, einer Energiefunktion, um Proteinmodelle zu bewerten und einem Optimierungsverfahren, um Proteinmodelle zu optimieren. Da TRANCENT auf dem Proteindesignkonzept aufbaut, werden diese Komponenten im folgenden Text detaillierter vorgestellt.

2.3.1.1 Modellierung

Beim Proteindesign wird eine optimale Sequenz für eine gegebene Struktur gesucht. Um zu beurteilen, wie gut Sequenzen zur Struktur passen, müssen Proteinmodelle mit atomistischen Details generiert werden. Diese Aufgabe wird von der Modellierungseinheit übernommen. Da das Rückgrat als starr angenommen wird, beschränkt sich das Modellieren beim Proteindesign auf das Platzieren der Aminosäureseitenketten. Diese können je nach Aminosäuretyp unterschiedliche Konformationen annehmen. Diese Flexibilität beruht vor allem auf den Rotationsmöglichkeiten der Atombindungen.

Um diese Flexibilität für die Optimierung beherrschbar zu machen, wird beim Proteindesign nur eine definierte Menge von möglichen Seitenkettenkonformationen betrachtet, die so genannten Rotamere (Dunbrack, 2002). Diese repräsentieren besonders wahrscheinliche Seitenkettenkonformationen. Die Rotamere sind in einer Rotamerbibliothek (Dunbrack, 2002; Ponder & Richards, 1987) zusammengefasst und werden in Form von Rotationswinkeln für alle drehbaren Bindungen beschrieben. Die Rotamerbibliothek wird aus einer Häufigkeitsanalyse von einzelnen Rotameren in bekannten Proteinstrukturen abgeleitet. Da die einzelnen Rotamere nicht gleich häufig beobachtet werden, wird in einer Rotamerbibliothek zu jedem Rotamer auch die relative Häufigkeit des Rotamers gespeichert (Dunbrack & Cohen, 1997).

Da die Seitenketten der Aminosäuren je nach Art eine unterschiedliche Anzahl rotierbarer Atombindungen besitzen, ist die Größe des Konformationsraums für Seitenketten aminosäurespezifisch (Abbildung 5).

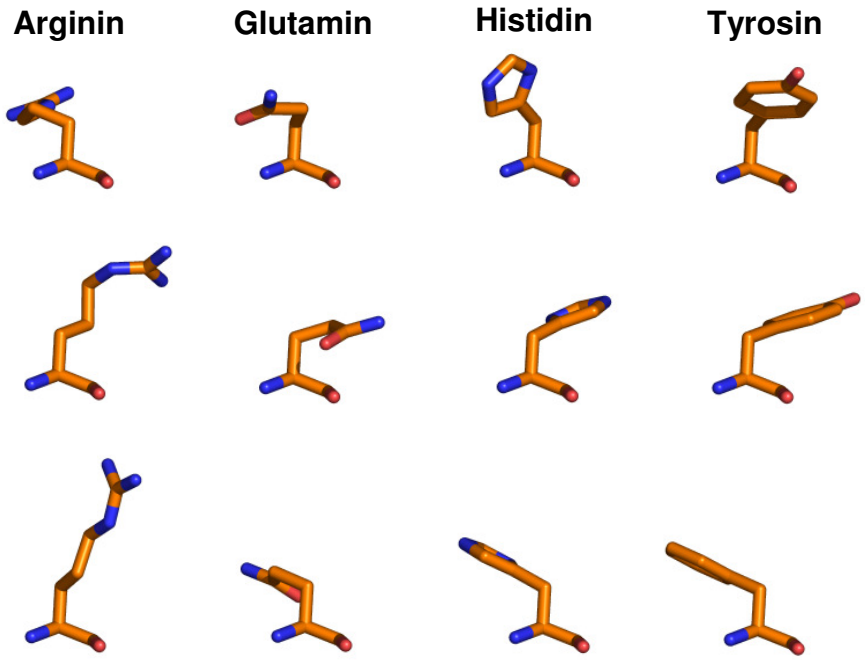


Abbildung 5: Beispiele für Rotamerausprägungen
Rotamere sind energetisch günstige Seitenkettenkonformationen von Aminosäuren. In der Abbildung sind für vier verschiedene Aminosäuren jeweils drei verschiedene Rotamere zu sehen. Arginin besitzt vier drehbare Bindungen mit jeweils drei besonders günstigen Winkeln. Daher ergeben sich für Arginin 81 Rotamere (3^4). Für Glutamin geben sich mit 3 drehbaren Bindungen 27 Rotamere. Tyrosin und Histidin haben mit nur jeweils zwei drehbaren Bindungen nur 9 Rotamere.

Entsprechend ist in der Rotamerbibliothek die Anzahl der Rotamere pro Aminosäure unterschiedlich. Glycin (Gly) und Alanin (Ala) besitzen keine rotierbaren Bindungen in den Seitenketten. Deswegen werden diese Aminosäuren durch jeweils ein Rotamer repräsentiert. Die Seitenketten von Arginin (Arg) und Lysin (Lys) sind dagegen lang gestreckt und flexibel. Mit vier rotierbaren Bindungen und drei energetisch günstigen Winkeln pro Bindung ergeben sich jeweils 81 Rotamere (3^4).

Die Menge der heute verfügbaren Proteinstrukturen erlaubt es, die Rotamerverteilungen positionsspezifisch (engl. *backbone dependent*) auszuwerten (Dunbrack & Karplus, 1993). Die Rotamere werden dazu in Abhängigkeit von den Φ- und Ψ-Winkeln des Rückgrats beschrieben (Ramachandran et al., 1963).

Die Modellierungseinheit platziert Aminosäureseitenketten auf Positionen im Proteinrückgrat und stellt dann für ein ausgewähltes Rotamer die entsprechenden

Bindungswinkel ein. Durch die Verwendung von Rotameren im Proteindesign kann es passieren, dass die Seitenketten nicht genau genug modelliert werden können, weil die benötigten Konformationen „zwischen“ den verfügbaren Rotameren liegen (vgl. Abbildung 6). Die Modellierungseinheiten einiger Proteindesignprogramme erlauben daher, die Menge der in der Rotamerbibliothek beschriebenen Rotamere durch solche mit leicht variierten Winkeln zu erweitern (De Maeyer et al., 1997). Dadurch wird der Konformationsraum der Seitenketten feiner aufgelöst und das Problem abgeschwächt. Allerdings wird auf diese Weise der Suchraum vergrößert.

Die Beschränkung der Seitenkettenkonformationen auf eine feste Anzahl verschiedener Rotamere dient dazu, die Komplexität bei der Suche nach einer optimalen Sequenz beherrschbar zu machen. Dem gleichen Zweck dient die Verwendung eines starren Rückgrats. Auch dabei handelt es sich um eine Approximation, weil optimale Interaktionen der Seitenketten subtile Änderungen der lokalen Rückgratstruktur erfordern können. Wenn solche Freiheitsgrade des Rückgrats während der Sequenzoptimierung (Desjarlais & Handel, 1999; Georgiev & Donald, 2007) berücksichtigt werden, wird die Modellierung genauer, die Optimierung jedoch aufwändiger.

Da bei der Energieberechnung auch die Wechselwirkungen der Wasserstoffatome berücksichtigt werden und diese in Proteinstrukturen oft nicht aufgelöst sind, muss die Modellierungseinheit dem Rückgrat und den generierten Seitenketten Wasserstoffatome hinzufügen können (Word et al., 1999).

2.3.1.2 Energiefunktion

Beim Proteindesign dient die Energiefunktion dem Zweck, die Stabilität von Proteinmodellen zu bewerten (Boas & Harbury, 2007; Gordon et al., 1999; Kuhlman & Baker, 2000; Liang & Grishin, 2004; Pokala & Handel, 2005). Die Energiefunktion modelliert dazu die physikochemischen Effekte der Stabilisierung. Die Qualität der Energiefunktion ist entscheidend für die Vorhersagegenauigkeit der Proteindesignmethode. Nur wenn die Energiefunktion die Stabilisierungseffekte richtig berechnet, bildet ein optimiertes Proteinmodell die Struktur des realen Proteins in hinreichender Genauigkeit ab.

2.3.1.2.1 Energieterme

Eine Energiefunktion setzt sich üblicherweise aus einer Menge einzelner Energieterme zusammen (Boas & Harbury, 2007). Typischerweise werden die einzelnen Effekte als separate Energieterme definiert. Beispiele sind Terme für die Van-der-Waals-Wechselwirkung, für Wasserstoffbrücken, die Ladungsinteraktion und für die

Solvatation. In der Energiefunktion werden sie als gewichtete Linearkombination zusammengeführt.

$$E_{gesamt} = w_1 E_{Van-Der-Waals} + w_2 E_{H-Brücken} + w_3 E_{Solvatation} + w_3 E_{Elektrostatik} \ldots$$

Durch die Gewichte (w_1, w_2 …) werden die Energieterme gegeneinander austariert (Kuhlman & Baker, 2000). Die Terme selbst und vor allem die dafür benötigten Parameter wie Ladung und Atomradien werden oft aus Kraftfeldern für die Moleküldynamik (Mackerell, 2004) übernommen.

Da beim Proteindesign das Rückgrat als starr angenommen wird und die Seitenketten als diskrete Rotamere modelliert werden, müssen die Energieterme an diese Approximation angepasst werden. Dies soll an einem Beispiel erläutert werden. Zur Modellierung der Van-der-Waals-Wechselwirkung wird in Kraftfeldern das klassische Lennard-Jones-Potential verwendet, das aus einem Anziehungsterm und einem Abstoßungsterm besteht:

$$E_{Lennard\ Jones}(r) = 4\varepsilon\left\{\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right\}$$

Die Funktion ist abhängig vom Abstand r zweier Atome und wird durch ε skaliert. Der Abstand, bei dem sich Anziehung und Abstoßung aufheben, ist durch σ gegeben. Energetisch optimal ist der Abstand $r_m = 2^{1/6}\sigma$. Da die Atome optimal angeordneter Seitenketten beim Rotamermodell überlappen können (Abbildung 6), würde der Abstoßungsterm, der in der 12. Potenz mit der Überlappung steigt, die Energie zu hoch bewerten. Bei der Optimierung würde diese Konfiguration zugunsten einer anderen ohne Überlappung verworfen. Um diesen Effekt zu kompensieren, werden zum Beispiel die Atomradien um bis zu 10% verkleinert und der Abstoßungsterm wird durch eine lineare Funktion ersetzt (Kuhlman & Baker, 2000).

Alternativ werden auch wissensbasierte Energieterme (Poole & Ranganathan, 2006; Sippl, 1995) für Energiefunktionen verwendet. Beim wissensbasierten Ansatz wird die Energie von Zuständen (z.B. Atomabstand oder Bindungswinkel) aus deren Häufigkeit in einer beobachteten Stichprobe abgeleitet. So nutzt das Proteindesign-Programm ROSETTA DESIGN (Kuhlman & Baker, 2000) kein physikalisch motiviertes Coulomb-Potential, sondern ein rein wissensbasiertes Paarpotential, um die Ladungsinteraktion zu modellieren. Das wissensbasierte Paarpotential wurde aus den Abständen von Aminosäurepaaren in bekannten Proteinstrukturen abgeleitet.

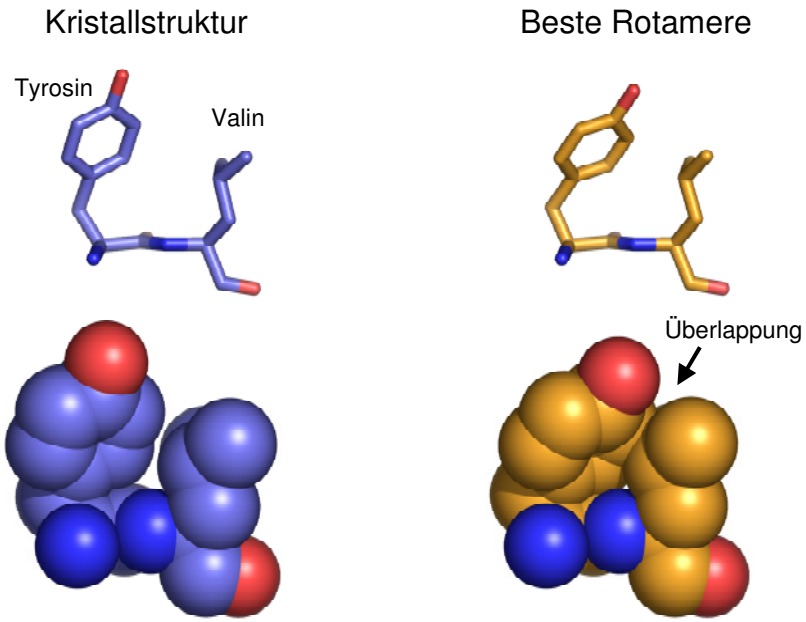


Abbildung 6: Überlappende Atome verursacht durch Rotamerapproximation
Links oben sind zwei Aminosäuren (Tyrosin und Valin) dargestellt, wie sie in der Kristallstruktur eines Proteins vorliegen. Rechts oben sind die Seitenketten dieser Aminosäuren durch die ähnlichsten Rotamere einer Rotamerbibliothek angenähert. Unten sind beide Fälle als Volumenmodell dargestellt. Der Pfeil deutet auf die Atome, die sich aufgrund der Rotamerapproximation durchdringen.

2.3.1.2.2 Paarweise zerlegbare Energiefunktionen

Eine Energiefunktion wird „paarweise zerlegbar“ genannt (Street & Mayo, 1998), wenn für alle Energieterme gilt, dass maximal zwei Aminosäurepositionen des Proteins gleichzeitig betrachtet werden müssen, um die Menge aller Teilenergien eines Proteinmodells zu berechnen. Da im Proteindesign die Modelle als fixes Rückgrat und eine Konfiguration von *n* Rotameren beschrieben werden, lässt sich die Energie solcher Energieterm als Summe der inneren Energien von Rückgrat und Rotameren, sowie aller Interaktionsenergien der Rotamere mit dem Rückgrat und untereinander (Dahiyat & Mayo, 1997) beschreiben:

$$E_{X,gesamt} = E_{X,Rückgrad} + \sum_{i=1}^{n} E_{X,Rotamer_i} + \sum_{i=1}^{n} E_{X,Rotamer_i,Rückgrad} + \sum_{i=1}^{n}\sum_{j=1}^{n} E_{X,Rotamer_i,Rotamer_j}$$

Hierbei steht E_x für einen der betrachteten Energieterme, die jeweils einen Beitrag zur Energie der paarweisen Energiefunktion liefern. Ist es für die Berechnung der Energie erforderlich, mehr als zwei Positionen simultan zu betrachten, spricht man von einer Multikörperenergiefunktion. Multikörperenergiefunktionen basieren zum Beispiel auf Solvatationstermen mit exakter Oberflächenberechnung (Leaver-Fay et al., 2007). Die

Oberflächenberechnung lässt sich bei paarweiser Zerlegung nur approximieren (Pokala & Handel, 2004; Street & Mayo, 1998). Die Eigenschaft der paarweisen Zerlegbarkeit ist eine wichtige Eigenschaft für die Auswahl des Optimierungsverfahrens. Es gibt zwar Verfahren, die es erlauben Multikörperenergiefunktionen zu optimieren, jedoch setzen die meisten Verfahren eine paarweise Zerlegbarkeit der Energiefunktion voraus (Siehe Kapitel 2.3.1.3).

2.3.1.2.3 Strafterme

Werden zusätzlich zur Stabilität weitere Bedingungen an die gesuchte Lösung gestellt, so kann die Energiefunktion um Strafterme erweitert werden. Die Energiefunktion beschreibt dann nicht nur die Stabilität des Proteinmodells, sondern auch, bis zu welchem Grad die zusätzlichen Bedingungen erfüllt worden sind.

In den Energiefunktionen molekülmechanischer Kraftfelder (Mackerell, 2004) sind solche Strafterme fester Bestandteil, um zum Beispiel Bindungslängen und Bindungswinkel optimal einzustellen. Dazu wird für einen Bindungswinkel oder eine Bindungslänge ein Referenzwert vorgegeben. Aus der Abweichung eines Modells von den Referenzwerten wird dann eine Strafenergie berechnet.

Auch bei der Strukturaufklärung mittels NMR (Wüthrich, 1995) werden Strafterme verwendet. Hier werden experimentell ermittelte Abstandsinformationen als Strafterme formuliert und der Energiefunktion eines Kraftfeldes hinzugefügt. Durch ein Energieminimierungsprotokoll lässt sich dann die gesuchte Struktur berechnen.

Im Proteindesign werden Strafterme verwendet, um z.B. Seitenketten in aktiven Zentren auszurichten oder um bestimmte Seitenkettenkonstellationen zu bevorzugen oder zu unterdrücken (Lassila et al., 2006).

2.3.1.3 Optimierungsverfahren

Die dritte Komponente beim Proteindesign bilden die Optimierungsverfahren (Desjarlais & Clarke, 1998; Voigt et al., 2000), die für die Suche nach einer Rotamerkonfiguration mit niedriger Energie zur Verfügung stehen. Die Rotamerkonfiguration mit der niedrigsten Energie (Desmet et al., 1992) wird GMEC (*global minimum energy conformation*) genannt. Verfahren die das GMEC berechnen, werden exakte Verfahren genannt. Dazu gehören *Dead Ends Elimination* (DEE) (Desmet et al., 1992; Goldstein, 1994) und *Branch and Terminate* (BT) (Gordon & Mayo, 1999).

Proteindesign gehört zur Klasse der NP-harten Problem (Pierce & Winfree, 2002). Für Proteine ab einer bestimmten Größe lassen sich Sequenzoptimierungen nicht mehr exakt lösen, da der Rechenaufwand exponentiell mit der Größe wächst.

In solchen Fällen wird auf heuristische Verfahren zurückgegriffen, wie zum Beispiel *Simulated Annealing* (SA) (Kirkpatrick et al., 1983; Metropolis et al., 1953) und Genetische Algorithmen (GA) (Holland, 1993), „*Self Consistent Mean Field*"-Optimierung (SCMF) (Koehl & Delarue, 1994; Lee, 1994) sowie „*Fast and accurate side-chain topology and energy refinement*" (FASTER) (Allen & Mayo, 2006; Desmet et al., 2002).

Heuristische Verfahren durchsuchen nicht den gesamten Lösungsraum, sondern beschränken sich auf Bereiche, die besonders vielversprechend sind. Daher kann nicht garantiert werden, dass die gefundene Lösung das GMEC ist. Bei manchen heuristischen Verfahren ist die Lösungssuche an einen Zufallsprozess gekoppelt (z.B. SA und GA), so dass bei verschieden Optimierungsläufen des gleichen Problems unterschiedliche Lösungen gefunden werden können. Wird bei solchen Verfahren unter Verwendung verschiedener Startbedingungen die beste Lösung mehrfach gefunden, so ist dies ein Indikator dafür, dass es sich bei der Lösung um die GMEC handeln könnte.

Das Finden der GMEC im Proteindesign garantiert nicht, dass die vorhergesagte Sequenz tatsächlich das stabilste Protein beschreibt, denn die Vorhersagequalität hängt auch von der Genauigkeit der Energiefunktion ab. Durch die Modellierungseinheit werden zusätzliche Ungenauigkeiten eingeführt, da diskrete Rotamere und ein starres Rückgrat verwendet werden. Es ist trotzdem vorteilhaft, exakte Optimierungsverfahren zu verwenden, da sich dadurch die Evaluierung vereinfacht. Wenn beim Proteindesign die Vorhersage vom Experiment abweicht, kommen als Fehlerquellen nur die Modellierungseinheit und die Energiefunktion in Frage, sofern exakte Optimierungsverfahren verwendet wurden.

Wird die Variabilität an jeder Position auf die möglichen Konformationen einer einzigen Aminosäure beschränkt, so ist das Proteindesignproblem auf einen Spezialfall reduziert (Lee & Subbiah, 1991; Summers & Karplus, 1989). In diesem Fall liefern die Optimierungsverfahren eine optimale Seitenkettenanordnung der gegebenen Sequenz. Anwendung findet die Seitenkettenoptimierung vor allem in der Homologiemodellierung (Chothia & Lesk, 1986; Ginalski, 2006). Die Seitenkettenoptimierung ist im Vergleich zum vollen Proteindesign deutlich weniger rechenaufwändig (Canutescu et al., 2003).

2.3.1.3.1 Simulated Annealing

Simulated Annealing (Kirkpatrick et al., 1983; Metropolis et al., 1953) ist ein heuristisches Verfahren, dessen Konzept an den physikalischen Abkühlungsprozess angelehnt ist. Während der Ausführung wird eine Startkonfiguration durch Zufallsschritte verändert und jeweils neu bewertet. Veränderungen werden zum Teil auch dann akzeptiert, wenn sich hierdurch die Bewertung der Konfiguration verschlechtert. Die

Entscheidung über die Annahme ungünstiger Änderungen ist dabei vom Zufall und von der Höhe einer virtuellen Temperatur abhängig, die im Verlauf der Optimierung sinkt.

Beim Proteindesign wird SA verwendet, um eine besonders energiearme Rotamerkonfiguration zu finden. Dazu wird mit einer Konfiguration von zufälligen Rotameren begonnen und deren Energie berechnet. Iterativ wird dann eine bestimmte Anzahl von Optimierungsschritten durchgeführt. In jedem Schritt wird zufällig eine Position gewählt, an der das vorhandene Rotamer ebenfalls zufallsgesteuert durch ein anderes ersetzt wird. Dieser Schritt führt entweder zu einer Änderung der Seitenkettenkonformation oder zum Wechsel der Aminosäure. Anschließend wird die Energie der neuen Konfiguration berechnet. Ist die Energie gesunken, wird die Änderung beibehalten und direkt zum nächsten Schritt übergegangen. Ist die Energie höher, wird mit Hilfe des Metropoliskriteriums (Metropolis et al., 1953) entschieden, ob der Schritt rückgängig gemacht oder akzeptiert wird. Dazu wird eine Akzeptanzschwelle p in Form einer Wahrscheinlichkeit bestimmt. Diese hängt von der berechneten Energiedifferenz und dem Abkühlungsgrad der virtuellen Temperatur T in folgender Weise ab:

$$p = \exp(\frac{1}{a^i T}(E_{neu} - E_{alt}))$$

Dabei ist a eine Abkühlungskonstante und i die Anzahl der bereits berechneten Schritte. Das Akzeptieren von energetisch ungünstigen Schritten dient dazu, lokale Minima der Energielandschaft zu überwinden. Durch Absenken der Temperatur im Laufe der Optimierung wird erreicht, dass die Akzeptanzwahrscheinlichkeit für Schritte, welche die Energiebilanz verschlechtern, stetig kleiner wird. Ein typisches SA Protokoll hebt im Verlauf der Optimierung die Temperatur mehrfach an und senkt sie dann wieder ab. Schließlich wird die Optimierung mit einer Quenchphase beendet. Beim Quenchen werden nur noch Rotameraustausche akzeptiert, welche die Energiebilanz verbessern.

SA findet Rotamerkonfigurationen mit niedriger Energie, es kann aber nicht garantiert werden, dass es sich dabei um die GMEC handelt. Mit großer Wahrscheinlichkeit handelt es sich jedoch um die bestmögliche Lösung, wenn diese bei Wiederholung der Optimierung mehrfach wieder gefunden wird und alle anderen Lösungen höhere Energie haben.

Es gibt verschiedene Gründe SA anstelle einer exakten Methode zu verwenden. SA wird vor allem dann genutzt, wenn die Problemgröße, also der Suchraum, für exakte Methoden zu groß wird. Im Gegensatz zur Optimierungsqualität ist die Laufzeit beim SA nicht von der Problemgröße abhängig. Durch mehrfache Wiederholung kann SA als stochastische Methode verschiedene niederenergetische Lösungen finden. Beim Proteindesign lassen die Gemeinsamkeiten und Unterschiede in den Sequenzen dann Rückschlüsse über die Bedeutung einzelner Sequenzpositionen für die Stabilität zu.

Außerdem fordert SA für die Energiefunktion keine paarweise Zerlegbarkeit, d.h. auch Multikörperenergiefunktionen lassen sich mit SA optimieren (Leaver-Fay et al., 2007).

2.3.1.3.2 Dead Ends Elimination

Dead Ends Elimination Optimierung basiert auf dem Konzept, diejenigen hochenergetischen Rotamere, die nicht Teil der GMEC sein können, zu identifizieren und auszuschließen (Desmet et al., 1992). Dazu wird iterativ ein Eliminierungskriterium auf die sich immer weiter verkleinernde Menge der noch nicht eliminierten Rotamere angewandt.

Für ein Protein mit N Positionen greift das Eliminierungskriterium in folgender Situation: Seien i_r und i_t zwei Rotamere an der Position i. Seien j_s Nachbarrotamere an allen anderen Positionen $j \neq i$. Wenn über alle Positionen j die Summe der kleinsten paarweisen Energiebeiträge für i_r mit j_s höher ist als die Summe der größten Energiebeiträge für i_t mit j_s, dann kann i_r nicht Teil der GMEC sein und darf eliminiert werden:

$$E(i_r) + \sum_{j \neq i}^{N} \min E(i_r, j_s) > E(i_t) + \sum_{j \neq i}^{N} \max E(i_t, j_s)$$

Anschaulich formuliert besagt die Ungleichung, dass ein Rotamer i_r eliminiert wird, wenn es trotz der besten Einbettung immer noch eine schlechtere Energiebilanz hat als ein alternatives Rotamer i_t mit der schlechtesten Einbettung.

Zu diesem Kriterium gibt es inzwischen eine Reihe von Erweiterungen, die mit der ursprünglichen Formulierung kombiniert werden können und so die Leistungsfähigkeit der DEE Methode erheblich verbessern (Georgiev & Donald, 2007; Goldstein, 1994; Lasters et al., 1995; Looger & Hellinga, 2001; Pierce et al., 2000).

Wenn die DEE-Optimierung konvergiert, wird die GMEC gefunden. Konvergieren bedeutet, dass an allen Positionen sämtliche Rotamere bis auf jeweils eines eliminiert werden konnten. Tritt die Konvergenz nicht ein, ist die verbleibende Menge der Rotamere nicht eindeutig und es sind immer noch mehrere Lösungen möglich. Für diesen Fall gibt es Ansätze, die Suche auf den verbleibenden Rotameren mit anderen Optimierungsmethoden fortzusetzen (Gordon & Mayo, 1999).

Mit DEE lassen sich nicht beliebig große Proteindesignprobleme optimieren. Die Laufzeit und die Konvergenzwahrscheinlichkeit sind stark an die Problemgröße gekoppelt. Da mit den Eliminierungskriterien immer Paare von Rotameren untersucht werden, muss die Energiefunktion paarweise zerlegbar sein. Für Probleme mit einer Multikörper-Energiefunktion ist DEE nicht anwendbar (Desmet et al., 1992).

2.3.1.4 Vorberechnung und Tabellierung

Eine wichtige Technik zum Beschleunigen des Proteindesignprozesses ist das konsequente Vorberechnen und Tabellieren aller relevanten Daten (Chowdry et al., 2007; Enkler, 2006). Durch einfaches Nachschlagen in Tabellen wird so eine zeitaufwändige Mehrfachberechnung vermieden. Dazu werden alle Rotamermodellierungen, sowie alle elementaren Energiebeiträge vorberechnet und tabelliert. Elementare Energiebeiträge sind die Summanden, die zur Gesamtenergie eines Proteinmodells zusammenaddiert werden (vgl. Kapitel 2.3.1.2). Durch die Tabellierung entsteht ein Datenmodell, das es erlaubt, die Strukturoptimierung durchzuführen, ohne auf die Modellierungseinheit oder die Energiefunktion zugreifen zu müssen. Stattdessen werden alle benötigten Informationen nur nachgeschlagen. Die Tabellen repräsentieren damit jeweils die Modellierungseinheit und die Energiefunktion. Diese Technik beschleunigt den Optimierungsprozess enorm.

Das vorab berechnete Datenmodell besteht aus zwei Tabellen. In der ersten Tabelle werden die Atomkoordinaten aller Rotamere gespeichert, die für die Modellierung relevant sein können. Hierfür werden an jeder Position des Proteinrückgrats die Seitenketten aller Aminosäuren angeordnet. Für jede Aminosäure werden iterativ alle Seitenkettenkonformationen, die in der Rotamerbibliothek beschrieben sind, eingestellt und die resultierende Seitenkettenkonformation in der Rotamertabelle abgelegt. Pro Aminosäureposition ergeben sich für die 20 Aminosäuren ca. 300 Rotamere als Einträge in der Rotamertabelle. Für das vollständige Design eines Proteins mit 100 Aminosäuren entsteht so eine Rotamertabelle mit 30.000 Rotameren.

Obwohl sowohl eine Rotamerbibliothek als auch eine Rotamertabelle einzelne Rotamere beschreiben, gibt es zwischen beiden Datenstrukturen wichtige Unterschiede. Die Rotamerbibliothek beschreibt die Rotamere abstrakt durch die Rotationswinkel und unabhängig von einem konkreten Proteinrückgrat. Sie liefert nur die Vorlage für die Rotamere der Rotamertabelle. In letzterer sind die Seitenketten als Menge von Atompositionen im kartesischen Raum relativ zu den assoziierten Positionen im Proteinrückgrat beschrieben. Daher können Wechselwirkungen zwischen den Rotameren der Rotamertabelle berechnet werden. Dies ist für die abstrahierten Rotamere der Rotamerbibliothek nicht möglich.

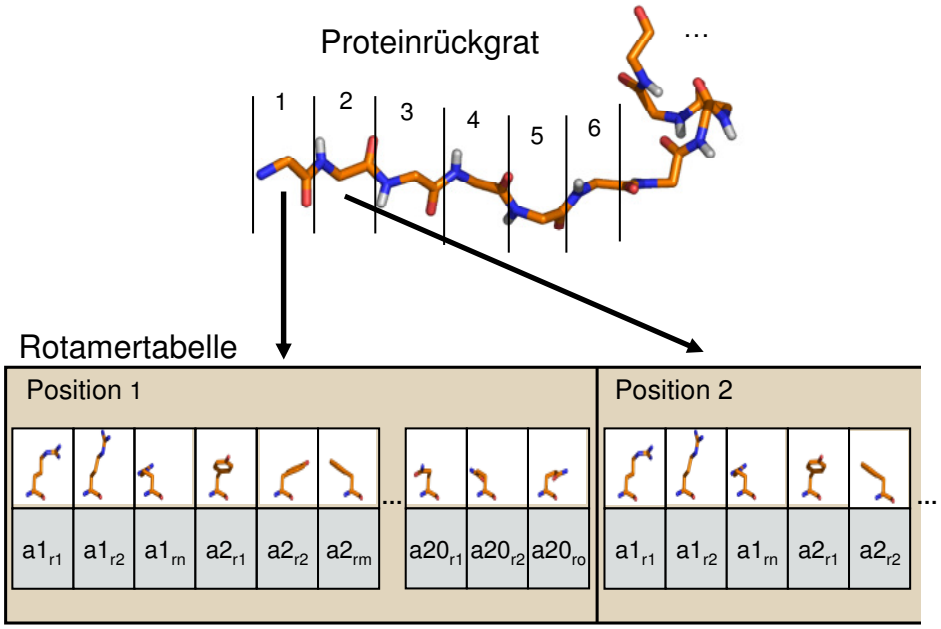


Abbildung 7: Schema einer Rotamertabelle
Die Menge der Rotamere, die an einer Position des Proteinrückgrats positioniert werden kann, ist beim Proteindesign endlich. Um einzelne Rotamere im Verlauf der Energieberechnung nicht immer wieder modellieren zu müssen, werden in einem Vorberechnungsschritt alle Rotamervarianten an jede Position modelliert und dann in einer Rotamertabelle abgespeichert. In der abgebildeten Tabelle werden die Atomkoordinaten der Rotamere positionsweise aufgelistet. Die Bezeichnung ai_{rj} für Einträge in die Tabelle bedeutet: Die i-te Aminosäure ($1 \leq i \leq 20$) in ihrem j-ten Rotamer ($1 \leq j \leq n$, n Anzahl der Rotamere der i-te Aminosäure).

In der zweiten Tabelle wird jeder Energiebeitrag für Paare von Rotameren, für Rotamere und mit dem Rückgrat, sowie für einzelne Rotamere selbst gespeichert (Abbildung 8). Dazu werden sukzessiv alle Rotamere und alle Rotamerpaare aus der Rotamertabelle an die Energiefunktion übergeben. Nachdem alle Energiebeiträge berechnet und gespeichert sind, lassen sich mit dieser Energietabelle die Gesamtenergien sämtlicher Strukturmodelle berechnen, die für das gegebene Strukturgerüst überhaupt modellierbar sind. Die geschieht durch einfaches Nachschlagen und Addieren der benötigten Energiebeiträge. Daher wird für das Optimierungsverfahren nur diese Energietabelle benötigt. Erst nach Abschluss der Optimierung, wird die Rotamertabelle noch einmal benötigt, um für die gefundene Lösung ein konkretes Strukturmodell zu generieren.

Für die Suche nach einer optimalen Lösung benötigen die Optimierungsverfahren im Idealfall nur die in der Energietabelle abgelegten Werte. Bei der Entwicklung eines Proteindesignverfahrens ist darauf zu achten, dass die Menge der Teilenergien in Bezug auf Rechenleistung und Speicherbedarf beherrschbar bleiben. Dies setzt voraus, dass die Energiefunktion paarweise zerlegbar ist. Ist die Energiefunktion eine Multikörper-

energiefunktion, so lässt sich die Energie nicht mehr vollständig in geeignet kleine Teilenergien zerlegen. Somit lassen sich nicht mehr alle Kombinationen mit vertretbarem Aufwand vorberechnen.

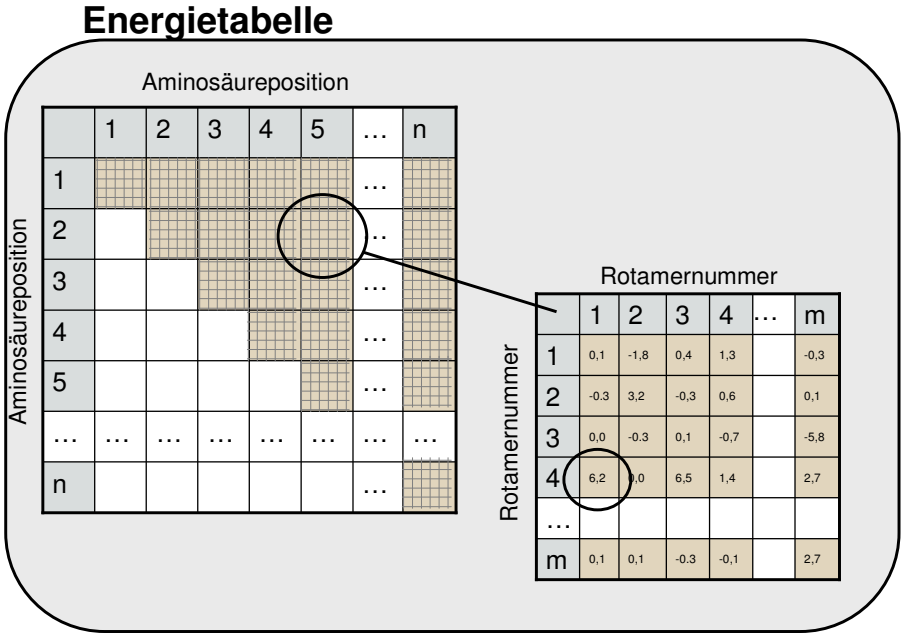


Abbildung 8: Schema einer Energietabelle
Um die Energieberechnung während der Optimierungsphase im Proteindesign zu beschleunigen, werden alle Teilenergien vorberechnet und in einer Tabelle abgelegt. Hier illustriert ist der Zugriff auf den Energiebeitrag der Interaktion zwischen dem vierten Rotamer an Position 2 und dem ersten Rotamer an Position 5 eines fiktiven Proteins. Die vorberechnete Interaktionsenergie beträgt hier 6.2.

Weiterhin ist zu beachten, dass der Speicherplatzbedarf auch bei einer paarweise zerlegbaren Energiefunktion quadratisch mit der Anzahl der Rotamere wächst. Für ein Proteindesignproblem mit 100 Positionen und 30.000 Rotameren ergibt sich der Speicherbedarf aus 450.000.000 Einträgen in der Energietabelle. Hierbei ist bereits berücksichtigt, dass die Energietabelle symmetrisch ist, so dass die Hälfte der Einträge nicht explizit gespeichert werden muss. Bei Speicherung der Teilenergien als Float-Werte (4 Byte) ergibt sich eine fast 2 Gigabyte große Tabelle. Da aber selbst langreichweitige Wechselwirkungen, wie elektrostatische Wechselwirkungen mit zunehmendem Abstand schwächer werden, gibt es Paare von Positionen, die so weit voneinander entfernt liegen, dass sich deren Rotamere energetisch nicht mehr nennenswert beeinflussen. Daher ist es üblich, Wechselwirkungen nur bis zu einem bestimmten Abstand (typisch 6Å) zu berücksichtigen. Deshalb bleiben bestimmte Bereiche der Energietabelle leer. Durch geschicktes Zusammenfassen dieser Regionen

lässt sich viel Speicherplatz sparen. Eine derart gestaltete Energietabelle erfordert für ein Proteindesignproblem mit 100 Positionen typischerweise weniger als ein Gigabyte an Speicherplatz (Chowdry et al., 2007; Enkler, 2006).

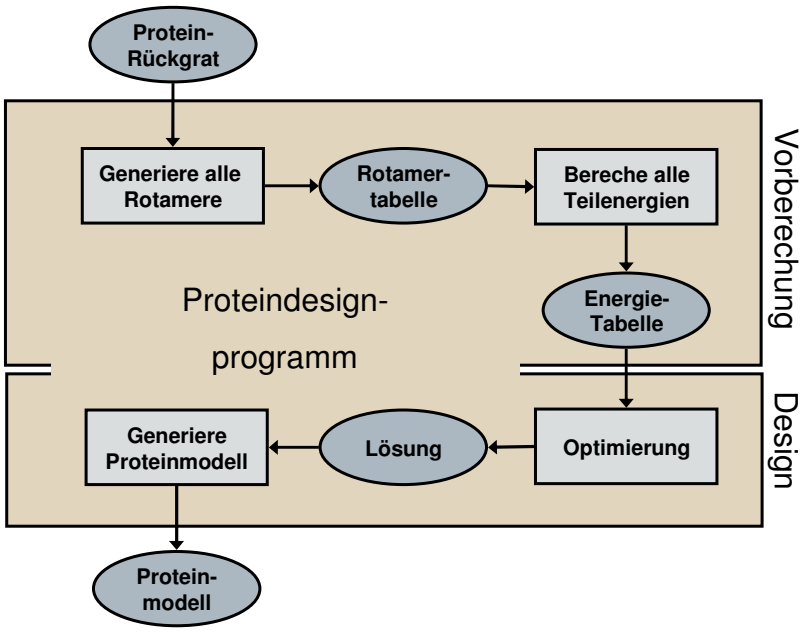


Abbildung 9: Flussdiagramm des Proteindesignprozesses
Der Prozess teilt sich in zwei Phasen: In der ersten Phase werden Daten vorberechnet. Hierfür werden alle erlaubten Rotamere des Proteinrückgrats modelliert. Dann werden alle Teilenergien für alle möglichen Rotamerkombinationen berechnet. Atomkoordinaten der Rotamere und Teilenergien werden in Tabellen gespeichert. Erst in der zweiten Phase läuft das eigentliche Proteindesign ab. Die Suche nach der optimalen Sequenz findet dabei ohne explizites Modellieren von Strukturen allein auf Basis der Energietabelle statt. Die gefundene Lösung ist eine Abfolge von Referenzen auf die Rotamertabelle. Im letzten Schritt werden die ausgewählten Rotamere schließlich zusammen mit dem Rückgrat zu einem Strukturmodell zusammengefügt.

Die beiden Proteindesignkomponenten Modellierungseinheit und Energiefunktion werden nach dem Generieren der Rotamertabelle und der Energietabelle nicht mehr benötigt. Die Tabellen repräsentieren die Komponenten, mit denen sie generiert wurden. Das Arbeiten mit den Tabellen ist nicht nur schneller, sondern auch technisch einfacher. Durch dieses Verfahren wird auch eine Entkopplung des Optimierungsverfahrens von der Komplexität der Energiefunktion erreicht.

2.3.1.5 Proteindesign-Programme

Erste Programme für das Proteindesign entstanden Anfang der neunziger Jahre (Hellinga & Richards, 1991). Seitdem profitiert das Feld nicht nur von der stetig

verbesserten Rechnerleistung und der wachsenden Menge an bekannten Proteinstrukturen, sondern in besonderem Maß auch von neuen methodischen Entwicklungen. Vier erfolgreiche Proteindesignprogramme, die den aktuellen Stand der Technik repräsentieren, werden nun genauer vorgestellt:

2.3.1.6 DEZYMER

DEZYMER ist ein Proteindesignprogramm aus der Arbeitsgruppe von Homme Hellinga (Hellinga & Richards, 1991; Looger & Hellinga, 2001). Mit diesem Programm lassen sich automatisch aktive Zentren von einem Protein auf ein anderes übertragen. Als DEZYMER im Jahre 1991 vorgestellt wurde, konnten bereits das Metallionen-bindende aktive Zentrum von Azurin auf Thioredoxin, ein Protein mit anderer Faltungstopologie und Funktion, übertragen werden. Mit einer verbesserten Version des Programms wurde 2004 das aktive Zentrum der Triosephosphatisomerase auf das Ribosebindeprotein übertragen (Dwyer et al., 2004). Dieses Resultat war der erste erfolgreiche Versuch, mit Hilfe eines Programms ein aktives Zentrum mit enzymatischer Funktion zu übertragen (vgl. Kapitel 2.2.3). Weiterhin konnte mit dem Programm für mehrere Enzyme die Bindung von neuen Liganden etabliert werden (Allert et al., 2004; de Lorimier et al., 2002; Looger et al., 2003). In allen genannten Fällen lässt sich die Bindung des neuen Liganden durch Änderung der Fluoreszenzeigenschaften nachweisen. Daher können diese neuen Proteine als Biosensoren für die jeweiligen Liganden verwendet werden.

Der Programmablauf ist bei DEZYMER in drei Phasen geteilt. 1) Die Suche nach geeigneten Stellen für das aktive Zentrum in der Struktur. 2) Die Optimierung der gefundenen Stellen. 3) Eine Sortierung der Lösungen entsprechend ihrer Qualität. In der ersten Phase werden nur das Rückgrat des Proteins und die für Katalyse oder Bindung essentiellen Reste betrachtet. Es werden systematisch alle Möglichkeiten getestet, die Seitenketten der essentiellen Reste so zueinander auf dem Rückgrat anzuordnen, dass zwischen ihnen der Ligand optimal ausgerichtet werden kann. Diese Konstellation von essentiellen Resten und Ligand wird zusammen mit dem Rückgrat für die nächste Phase starr fixiert und es werden die Seitenketten der übrigen Positionen des Proteins wieder hinzugefügt. In der zweiten Phase werden die Reste der Umgebung des Liganden durch eine Proteindesignroutine optimiert. Da die katalytischen Reste und der Ligand starr in ihrer Lage bleiben, optimiert die Routine deren Einbettung durch Reste, die energetisch besonders verträglich sind. Auf diese Weise wird das neue aktive Zentrum in seiner Gesamtheit stabilisiert. Schließlich werden in der dritten Phase die Lösungen entsprechend ihrer Energie sortiert.

Die Energiefunktion (Looger & Hellinga, 2001) von DEZYMER beruht auf den Parametern des CHARMM Kraftfeldes (MacKerell et al., 1998) und besteht aus Van-der-Waals-Term, einem expliziten Wasserstoffbrückenterm, einem Solvatationsterm

(Street & Mayo, 1998) der durch eine paarweise Zerlegung approximiert ist, einem Entropieterm sowie einem wissensbasierten Term, der sich aus statistisch ermittelten Aminosäureverteilungen ableitet.

Als Optimierungsverfahren wird DEE eingesetzt. Dies ist möglich, da die verwendete Energiefunktion paarweise zerlegbar ist. Durch die Beschränkung der Sequenzoptimierung auf das aktive Zentrum bleibt die Problemgröße mit DEE handhabbar.

2.3.1.7 ROSETTA DESIGN

ROSETTA DESIGN (Kuhlman & Baker, 2000) ist ein Proteindesignprogramm aus der Arbeitsgruppe von David Baker. Es ist Teil einer ganzen Reihe von Programmen zur Proteinmodellierung, die zum ROSETTA-Softwarepaket zusammengefasst sind. Für sich betrachtet ist ROSETTA DESIGN ein klassischer Ansatz, um für ein gegebenes Proteinrückgrat eine optimale Sequenz zu finden.

Es wurde gezeigt, das ROSETTA DESIGN in der Lage ist, zu einem vorgegebenen Proteinrückgrat einen Teil der wildtypischen Sequenz wieder zu finden. Im Proteininneren liegt dieser Anteil bei 51% der Reste, an der Oberfläche bei 27%. Diese Werte wurden mit einem Testdatensatz von etwa 100 Strukturen ermittelt. Dass durch ROSETTA DESIGN gefundene Sequenzen tatsächlich stabiler als die wildtypischen Sequenzen sind, konnte eindrucksvoll an mehreren Proteinen aus bis zu 70 Aminosäuren gezeigt werden (Dantas et al., 2007; Dantas et al., 2003).

Mit Hilfe von ROSETTA DESIGN konnte das erste Protein konstruiert werden, das einen in der Natur nicht vorkommenden Faltungstyp aufweist (Kuhlman et al., 2003). Das entsprechende Rückgrat wurde dazu vorgegeben. Das Programm wurde dann iterativ im Wechsel mit einer Rückgratoptimierung angewendet, bis die Lösung gegen eine stabile Struktur und Sequenz konvergierte. Das resultierende Protein konnte anschließend hergestellt und seine Struktur aufgeklärt werden. Die Struktur des Modells wich von der Kristallstruktur nur um 1.2 Å (Rückgrat-RMSD) ab.

Die in ROSETTA DESIGN verwendete Energiefunktion (Kuhlman & Baker, 2000) ist mit dem CHARMM Kraftfeld (MacKerell et al., 1998) parametrisiert und besteht aus einem van-der-Waals-Term, einem Wasserstoffbrückenterm, einem Solvatationsterm, einem wissensbasierten Paar-Potential-Term, der die Elektrostatik approximiert und einem wissensbasierten Selbstenergieterm, der sich aus den Wahrscheinlichkeiten für die verwendeten Rotamere ableitet. ROSETTA DESIGN verwendet SA zur Sequenzoptimierung.

2.3.1.8 HERO

HERO ist ein Proteindesignprogramm (Gordon et al., 2003) aus der Arbeitsgruppe von Stephen Mayo. Es stellt zusammen mit anderen Programmen der Arbeitsgruppe wie ORBIT (Dahiyat & Mayo, 1997) oder VEGAS (Shah et al., 2004) einen Teil eines größeren Programmpakets für das Proteindesign dar.

Mit dem Programm wurde zum Beispiel die komplette Sequenz aus 51 Aminosäuren des gekerbten Homeodomänenproteins aus *Drosophila melanogaster* auf Stabilität optimiert (Shah et al., 2007). Das neue Protein wurde synthetisiert und ist auch im Experiment deutlich thermostabiler als der Wildtyp.

Die Energiefunktion von HERO basiert auf dem DREIDING Kraftfeld (Mayo et al., 1990) und besteht aus vier Energietermen: Einem van-der-Waals-Term, einem Wasserstoffbrückenterm, einem Elektrostatik-Term und einem Solvatationsterm. Im Gegensatz zu ROSETTA DESIGN verwendet die Energiefunktion von HERO keine wissensbasierten Paar-Potentiale für die Ladungsinteraktion, sondern einen physikalisch motivierten Elektrostatik-Term (Marshall et al., 2005; Zollars et al., 2006).

Als Optimierungsverfahren verwendet HERO eine DEE Variante. Dabei wird DEE mit einem stochastischen Element verknüpft. Das Element wird verwendet, um eine möglichst optimale Reihenfolge zu finden, in der Rotamere eliminiert werden. Auf dieses Weise wird die Laufzeit zum Finden der optimalen Lösung verbessert, wobei im Konvergenzfall dennoch das GMEC gefunden wird (Gordon et al., 2003).

2.3.1.9 EGAD

EGAD (Pokala & Handel, 2004; Pokala & Handel, 2005) ist ein Proteindesign-programm aus der Arbeitsgruppe von Tracy Handel. Das Programm lässt sich nicht nur zur Proteinstabilisierung, sondern auch für eine ganze Reihe von weiteren Anwendungen nutzen. So können mit EGAD Ligandenbindungen optimiert oder pKa-Werte berechnet werden. Der Quellcode des Programms ist in C++ geschrieben und frei verfügbar. Zudem ist EGAD in Form der Programmierbibliothek EGAD LIB (Chowdry et al., 2007) nutzbar.

Die Energiefunktion in EGAD verwendet im Gegensatz zu anderen Energiefunktionen nur wenige empirische Parameter und beschreibt die Effekte durch physikalisch motivierte Terme. Die Energiefunktion basiert auf dem OPLS-AA Kraftfeld (Jorgensen et al., 1996), wobei eine eigene Implementation für die Elektrostatik und die Solvatation verwendet wird. Die Energiefunktion ist paarweise zerlegbar und daher bei der Auswahl des Optimierungsverfahrens nicht beschränkt. In der Programmierbibliothek sind praktisch alle der im Proteindesign verwendeten Optimierungsverfahren (SA, GA, SCMF, DEE, FASTER) implementiert.

2.3.2 Ligandenbindung

Neben der Proteinstabilität ist die Etablierung der Ligandenbindung eine wichtige Voraussetzung, um erfolgreich den Transfer von aktiven Zentren zu modellieren. Ohne Bindung ist keine Katalyse möglich. Um eine Bindestelle zu modellieren, muss zunächst eine Stelle im Protein gefunden werden, die den Liganden aufnehmen kann.

Abhängig von der Flexibilität der Ligandenstruktur ist es unter Umständen notwendig, verschiedene Konformationen zu berücksichtigen. Wenn der Ligand in einer geeigneten Bindungspose positioniert ist, bedarf es energetisch günstiger Wechselwirkungen zwischen dem Liganden und den Atomen der Bindetasche. Dazu müssen die Seitenketten so optimiert werden, dass der Ligand optimal eingebettet wird. Die Energiefunktionen der Proteindesignmethoden müssen für die Ligandenbindung angepasst sein, damit diese Wechselwirkungen beschrieben werden können. Insgesamt sind die Probleme bei der Ligandenbindung im Proteindesign den Problemen beim Liganden-Docking (Hirayama, 2007) und beim Wirkstoffdesign (Böhm et al., 1996) sehr ähnlich. Erschwerend kommt jedoch beim Proteindesign hinzu, dass neben der Ligandenbindung auch die Sequenz optimiert werden muss.

2.3.2.1 Ligandenpositionierung im Proteindesign

Ein vergleichsweise einfacher Weg, eine Bindestelle für den Liganden zu suchen, ist in (Bolon & Mayo, 2001) vorgestellt. Da hier die Katalyse von nur einem Rest, einem Histidin, abhängt, wird eine Hybridaminosäure (bestehend aus Histidin mit gebundenem Liganden) definiert. Anschließend werden Rotamere berechnet, welche die Freiheitsgrade des Liganden und des Histidins in Kombination abdecken. Diese Hybridrotamere werden anschließend zusammen mit der Menge der Standardrotamere im Proteindesign verwendet. Mit DEZYMER (Hellinga & Richards, 1991) lässt sich der Ligand zusammen mit mehreren Resten positionieren. Dazu wird das Proteinrückgrat systematisch nach einer Menge von Positionen durchsucht, in denen sich alle relevanten Rotamere in der richtigen Orientierung relativ zu einem Liganden anordnen lassen. In (Zanghellini et al., 2006) wird das Konzept um eine Hashing-Technik erweitert, um mehrfache Modellierung von Rotameren zu vermeiden und um sehr schnell entscheiden zu können, ob es zu bereits platzierten Rotameren weitere kompatible gibt. Außerdem wird in der gleichen Arbeit ein Ansatz („*inverse rotamer tree*“) vorgestellt, der es erlaubt, zunächst alle möglichen Ligand-Rotamer-Konfigurationen unabhängig von einem gegebenen Rückgrat zu berechnen. Einzelne Konfigurationen werden dann als Abstandsbedingungen für die Hauptkettenatome gespeichert. Anschließend wird ein gegebenes Rückgrat nach Positionskombinationen abgesucht, die Abstandsbedingungen einer der gespeicherten Konfigurationen erfüllen. Alle Ansätze fixieren im nächsten

Schritt die gefundenen Ausrichtungen der katalytischen Reste und des Liganden. Dann erst werden die übrigen Reste mit einer Proteindesignmethode optimiert, so dass die katalytisch relevanten Reste zusammen mit dem Liganden optimal eingebettet sind. Positionierung des Liganden und das eigentliche Proteindesign finden also separat statt.

Der Ansatz von (Zollars et al., 2006) kombiniert die Platzierung der katalytisch essentiellen Reste mit der Sequenzoptimierung, indem Rotamere die für die Katalyse besonders geeignet sind, durch einen Energiebonus bei der Optimierung bevorzugt werden.

2.3.2.2 Optimierung der Bindungsaffinität im Proteindesign

Um die Bindungsenergie zum Liganden zu optimieren, verwenden alle Proteindesignmethoden Varianten der Energiefunktion, mit der die Seitenketten optimiert werden (vgl. Kapitel 2.3.1.5). In (Meiler & Baker, 2006) wird beschrieben, wie eine Energiefunktion für das Programm ROSETTA DESIGN angepasst wurde.

Bei kraftfeldbasierten Energiefunktionen für Liganden ist die Parametrisierung kritisch. Parametrisieren bedeutet hierbei, für die einzelnen Atome eines Liganden die Bindungen, Radien, Protonierungen und Ladungen zu bestimmen. In Proteinen kommen maximal 20 unterschiedliche Aminosäuren vor. Daher ist die Parametrisierung für Proteine auf wenige Fälle beschränkt. Im Gegensatz zu Aminosäuren sind Liganden sehr vielfältig. Daher müssen einzelne Liganden individuell parametrisiert werden. Auch wenn inzwischen Programme dabei helfen können, diesen aufwändigen Prozess zu automatisieren (Schuttelkopf & van Aalten, 2004), bleibt die Ligandenparametrisierung ein kritischer Schritt. Denn nur bei geeigneter Parametrisierung werden die Wechselwirkungen zwischen Ligand und Protein vernünftig beschrieben.

2.3.2.3 DRUGSCORE

Das Programm DRUGSCORE (Gohlke et al., 2000) wurde in der Arbeitsgruppe von Gerd Klebe entwickelt, um Bindungsaffinitäten für Protein-Ligand-Komplexe zu berechnen. Es wird daher für das Docken von Liganden (Sousa et al., 2006; Wang et al., 2003) verwendet.

Die Scoringfunktion von DRUGSCORE ist wissensbasiert. Dazu wurden bekannte Protein-Ligand-Komplexstrukturen aus der RELIBASE Datenbank (Hendlich, 1998) analysiert, um Abstandsverteilungen von Protein-Ligand-Atompaaren abzuleiten. Zusätzlich berücksichtigt DRUGSCORE auch den Verlust an Lösungsmittelzugänglichkeit von Atomen, die an der Komplex-Interaktion beteiligt sind. Hier wurde für jeden Atomtyp eine Verteilung aus bekannten Komplexen ermittelt, die beschreibt, zu welchem Grad die Komplexbildung die Lösungsmittelzugänglichkeit verringert. Aus

den Verteilungen wurden dann wissensbasierte Potentiale abgeleitet (Sippl, 1993) und in einer Scoringfunktion zusammengefasst. Um mit der Scoringfunktion einen Protein-Ligand-Komplex zu bewerten, wird berechnet, wie gut die Abstände der Atompaare des zu bewertenden Komplexes den beobachteten Abstandverteilungen entsprechen.

Die maximale Distanz der Atompaare, die in den Verteilungen berücksichtigt werden, liegt bei 6Å. Bei größerem Abstand werden die langreichweitigen elektrostatischen Wechselwirkungen immer schwächer, so dass sie vernachlässigt werden können. Der Abstand von 6Å ist aber ausreichend, um kontaktvermittelnde implizite Wassermoleküle zu berücksichtigen. Solche impliziten Wassermoleküle liegen zwischen Ligand und Protein, sind aber nicht in den Strukturdaten beschrieben.

Da die Ableitung der Potentiale nur schwere Atome (keinen Wasserstoff) berücksichtigt, ist eine Modellierung der Protonen von Protein und Ligand nicht notwendig. Ein weiterer Vorteil des wissensbasierten Ansatzes ist, dass die Liganden nicht parametrisiert werden müssen.

Es gibt zwei Varianten des Programms, DRUGSCORE und DRUGSCORE CSD (Velec et al., 2005). Während die Beispielkomplexe für DRUGSCORE aus der RELIBASE Datenbank stammen, kommen die Beispiele für DRUGSCORE CSD aus der CSD Datenbank (Allen, 2002). Im direkten Vergleich ist DRUGSCORE CSD die genauere Variante.

2.3.3 Ähnlichkeit der aktiven Zentren

Der Transfer eines aktiven Zentrums bedeutet, die enzymatische Funktion, d. h. den Katalysemechanismus und die Bindestelle zu transferieren. Bei aktuellen Programmen werden diese beiden Aspekte getrennt. Reste, die als katalytisch essentiell klassifiziert sind, werden zusammen mit dem Liganden positioniert, so dass sie in der richtigen Geometrie vorliegen (vgl. Kapitel 2.3.2.1). Alle anderen Reste des aktiven Zentrums, das als Vorlage dient, werden als bindungsvermittelnd eingestuft und ignoriert. Anschließend wird die Wechselwirkung zum Liganden uniform, das heißt ohne Berücksichtigung der Vorlage, optimiert (vgl. Kapitel 2.3.2.2). Auf diese Weise kann der Transfer weitestgehend unabhängig von der Vorlage auch zwischen sehr verschiedenen Proteingerüsten durchgeführt werden (Dwyer et al., 2004). Im Extremfall beschränkt sich dann aber die Ähnlichkeit des neuen aktiven Zentrums zur Vorlage nur auf die katalytisch essentiellen Reste.

Da in der vorliegenden Arbeit aktive Zentren zwischen zwei Enzymen mit gleichem Faltungstyp getauscht werden sollen, ist Ähnlichkeit ein Indikator für eine besonders plausible Modellierung der Übertragung. Aus der Hypothese der divergenten Evolution folgt, dass beide Enzyme aus einem gemeinsamen Vorläufer hervorgingen. Daher

besitzen sie ähnliche strukturelle Voraussetzungen und auf beiden Strukturgerüsten sollte sich die jeweils andere Funktion auf ähnliche Weise etablieren lassen.

In den folgenden Kapiteln wird das Konzept der wissensbasierten Potentiale erläutert und es werden Methoden vorgestellt, die mit diesen Potentialen Ähnlichkeit optimieren.

2.3.3.1 Wissensbasierte Potentiale

Ein sehr vielseitig verwendbares Konzept, um Wissen aus Beispielen abzuleiten, sind wissensbasierte Potentiale. Diese Potentiale bewerten, wie gut eine beobachtete Merkmalsausprägung einem spezifischen Merkmal entspricht. Um solche Potentiale abzuleiten, werden zwei Stichproben benötigt: Eine mit spezifischen Beispielen des Merkmals zum Schätzen der Wahrscheinlichkeitsdichten $f_{spezifisch}$ und eine mit unspezifischen Beispielen für die Wahrscheinlichkeitsdichten $f_{unspezifisch}$. Aus den beiden Wahrscheinlichkeitsdichten wird dann über folgenden Ansatz das Potential definiert:

$$\Delta E(x) = -\ln\left(\frac{f_{spezifisch}(x)}{f_{unspezifisch}(x)}\right)$$

Hierbei ist x eine zu beurteilende Merkmalsausprägung. Dieser Ansatz lässt sich sowohl durch die Testtheorie und das Neyman-Pearson-Lemma (Merkl & Waack, 2002), als auch durch die physikalische Statistik und Invertierung des Boltzmann-Gesetz motivieren (Sippl, 1990). In diesem Fall wird die Formel noch um den Faktor kT für die Boltzmannkonstante k und die Temperatur T ergänzt.

Aus der Menge der bekannten Proteinstrukturen lassen sich wissensbasierte Potentiale ableiten. Dabei wird davon ausgegangen, dass sich die physikochemischen Wechselwirkungen zwischen den Atomen der Proteinstrukturen durch Abstandsverteilungen beschreiben lassen, die aus bekannten Strukturen abgeleitet sind (deduktives Prinzip).

In (Sippl, 1990) wurden auf diese Weise C_α-C_α-Abstandspotentiale für spezifische Aminosäurekombinationen *(a, b)* mit einem Sequenzabstand s bestimmt. Spezifische und unspezifische C_α- C_α-Abstände wurden in bekannten Proteinstrukturen vermessen und dienten dazu, Wahrscheinlichkeitsdichten abzuleiten. Mit den Wahrscheinlichkeitsdichten wurden anschließend Potentiale definiert:

$$\Delta E(a,b,s,x) = -kT\ln\left(\frac{f_{a,b,s}(x)}{f_s(x)}\right)$$

Hierbei steht x für den zu beurteilenden Abstand zwischen zwei Aminosäuren a und b mit dem Sequenzabstand s. Diese Potentiale wurden dann zum Beispiel für Threadingverfahren verwendet (Sippl & Flockner, 1996).

Weitere Beispiele für die Verwendung von wissensbasierten Potentialen sind die BLOSUM-Substitutionsmatrizen (Henikoff & Henikoff, 1992), DRUGSCORE (Gohlke et al., 2000) und das Programm MODELLER (Sali & Blundell, 1993).

2.3.3.2 Homologiemodelle

Um zu beurteilen, welche Strukturdetails in einem aktiven Zentrum von besonderer Bedeutung sind, können bekannte Strukturen des aktiven Zentrums auf Gemeinsamkeiten und Unterschiede hin untersucht werden. Diese Zusammenhänge lassen sich dann als Funktionsdefinition für das aktive Zentrum zusammenfassen. Für eine gesicherte Beurteilung ist die Größe der Beispielmenge entscheidend. Die PDB-Datenbank (Bernstein et al., 1977), in der bekannte Proteinstrukturen verwaltet werden, umfasst bereits mehr als 40.000 Strukturen und befindet sich in rasantem Wachstum. Trotzdem finden sich für einzelne Proteine oft nur eine oder gar keine Struktur und selten mehr als zehn. Andererseits gibt es für einzelne Proteine oft eine große Menge bekannter Proteinsequenzen. Nicht selten gibt es für ein Protein mehr als 500 homologe Sequenzen. Diese sind Vertreter des gleichen Proteins aus verschieden Organismen. Es liegt also nahe, Informationen aus diesen Sequenzen in die Funktionsdefinition einfließen zu lassen.

Eine Möglichkeit diese Informationen zu nutzen besteht in der Verwendung eines Multiplen Sequenzalignments (MSA). Für das MSA lassen sich unter Verwendung einer Beispielstruktur die Spalten ermitteln, die Reste des aktiven Zentrums repräsentieren. Die Beurteilung der Konserviertheit ermöglicht dann, die Bedeutung dieser Reste abzuschätzen. Der Ansatz ist aber unzureichend, wenn nicht nur die Konserviertheit sondern auch die strukturelle Variabilität erfasst werden soll. Ist zu den Sequenzen wenigstens eine Beispielstruktur vorhanden, so lassen sich aus den Sequenzen auch Homologiemodelle erzeugen (Ginalski, 2006). Die auf diese Weise generierten Modelle erweitern die Menge der Strukturbeispiele und erlauben die Ableitung relevanter Strukturdetails auf Grundlage einer größeren Datenbasis.

Bei der Homologiemodellierung soll für eine Sequenz eine Struktur vorhergesagt werden, wobei bereits eine Struktur mit ähnlicher Sequenz verfügbar ist. Es ist bekannt dass ein gewisser Grad an Sequenzähnlichkeit eine ähnliche Struktur impliziert (Sander & Schneider, 1991). Daher werden im Rahmen der Homologiemodellierung in Sequenzalignments Abschnitte bestimmt, die gut übereinstimmen. Für diese Bereiche wird die Struktur der Vorlage übernommen und an die neue Sequenz angepasst, indem die Seitenketten entsprechend ausgetauscht werden. Anschließend werden Bereiche, die sich nicht entsprechen (Insertionen, Deletionen) modelliert. Zuletzt werden die Wechselwirkungen im finalen Modell energetisch optimiert. Hierzu werden zum Beispiel die Seitenketten neu gepackt (vgl. Kapitel 2.3.1.3). Es bleibt festzuhalten, dass

die Genauigkeit des resultierenden Modells vor allem von der Sequenzähnlichkeit zur Vorlage abhängt.

Die Genauigkeit eines Homologiemodells lässt sich auch erhöhen, wenn mehrere ähnliche Vorlagen verwendet werden können. In dieser Situation lassen sich Gemeinsamkeiten der Vorlagen als konservierte Strukturdetails bestimmen und gegebenenfalls in das Modell übertragen. Im Programm MODELLER (Sali & Blundell, 1993) wird diese Idee mit so genannten MOLPDFs umgesetzt. Ein MOLPDF ist ein wissensbasiertes Potential, das als Bedingung in die Optimierung des Modells einfließt. Hierfür werden aus den Vorlagen für Strukturmerkmale Verteilungen geschätzt. Derartige Strukturmerkmale können Abstände oder Winkel zwischen Atomen benachbarter Seitenketten sein. Anschließend werden aus den Verteilungen wissensbasierte Potentiale abgeleitet. Diese Potentiale dienen dazu, ein Kraftfeld zu definieren, mit dem die Homologiemodelle optimiert werden können.

2.3.4 Katalyse und pKa-Werte

Bei Proteinen haben manche Seitenketten, sowie der N-Terminus und der C-Terminus titrierbare Gruppen. Titrierbar bedeutet, dass diese Gruppen unter bestimmten pH-Bedingungen ionisiert vorliegen, indem sie ein Proton abgeben oder aufnehmen (Voet & Voet, 2002). Bei Aspartat und Glutamat ist die Carboxylgruppe, beim Histidin der Imidazolring, beim Cystein das Thiol, beim Tyrosin die Hydroxylgruppe, beim Arginin die Guadiniumgruppe und beim Lysin die Amingruppe titrierbar. Die Protonierbarkeit dieser Gruppen wird durch den pKa-Wert beschrieben. Dies ist der pH-Wert, bei der die Gruppe mit gleicher Wahrscheinlichkeit protoniert oder deprotoniert vorliegen kann. Die titrierbaren Gruppen der Aminosäuren haben Referenz-pKa-Werte (Tabelle 2). Diese gelten für die Gruppen aber nur, wenn die Aminosäuren frei in wässriger Lösung vorliegen. Sind die Aminosäuren Teil eines Proteins, können sich deren pKa-Werte ändern. Die pKa-Werte werden dann als perturbierte Werte bezeichnet.

pKa-Perturbationen können durch verschiedene Effekte verursacht sein. Darunter fallen vor allem Lösungsmittelabschirmung, Wasserstoffbrücken und Ladungsinteraktionen. Wird eine titrierbare Gruppe durch die Proteinumgebung vom Lösungsmittel abgeschirmt und befindet sie sich dadurch in einer hydrophoben Umgebung, so ist die ungeladene Form der Gruppe stabiler und der pKa der Gruppe ist in Richtung der neutralen Form verschoben (Säuren zu höheren Werten hin, Basen zu niedrigeren Werten). Bilden titrierbare Gruppen Wasserstoffbrücken aus, so verschiebt sich deren pKa-Wert zu höheren Werten, falls sie Protonendonor sind und entsprechend zu niedrigeren, falls sie Protonenaktzeptor sind.

Tabelle 2: Referenz-pKa-Werte für Aminosäuren
In der Tabelle sind für die titrierbaren Gruppen einzelner Aminosäuren die Referenz-pKa-Werte aufgelistet. Diese Werte gelten für die einzelnen Aminosäuren, wenn sie frei in wässriger Lösung vorliegen.

Titrierbare Gruppe	Referenz pKa-Wert
C-Terminus	3.2
Aspartat	3.8
Glutamat	4.5
Histidin	6.5
N-Terminus	8.0
Cystein	9.0
Tyrosin	10.0
Lysin	10.5
Arginin	12.5

Für Ladungswechselwirkungen gilt folgendes: Falls sich die Ladungen abstoßen, verschieben sich die pKa-Werte in Richtung der neutralen Form. Ziehen sich die Ladungen an, so verschieben sie sich in entgegengesetzter Richtung. Bei der Ladungsinteraktion ist der Perturbationseffekt nicht nur vom Abstand der Ladungen, sondern auch sehr stark von deren Lösungsmittelzugänglichkeit abhängig. Da Wasser polar ist, schirmt es Ladungen an der Oberfläche des Proteins stark voneinander ab, der Perturbationseffekt ist gering. Im apolaren Proteininneren hingegen ist die Abschirmung der Ladungen viel geringer. Entsprechend stark sind mögliche Perturbationen der pKa-Werte durch Ladungsinteraktion. Da sich titrierbare Gruppen gegenseitig beeinflussen können, ist das Zusammenspiel dieser Effekte sehr komplex. So hängt beispielsweise der Ladungszustand einer titrierbaren Gruppe vom Ladungszustand aller anderen titrierbaren Gruppen in der Umgebung ab und umgekehrt.

Durch das komplexe Zusammenspiel dieser Wechselwirkungen ergeben sich für einzelne Gruppen unterschiedlich perturbierte pKa-Werte. Besonders häufig sind titrierbare Gruppen in aktiven Zentren stark perturbiert. Erst durch diese Perturbationen sind manche essentiellen Reste zur Katalyse unter physiologischen Bedingungen fähig. Solche Katalysen sind die Säure-Base-Katalyse oder die nukleophile Katalyse. Damit zum Beispiel ein Lysin-Rest als Nukleophil fungieren kann, muss er neutral vorliegen. Dazu muss der pKa-Wert der Amingruppe von 10.5 auf unter 7 abgesenkt sein. Für die Säurekatalyse muss ein aktiver Rest zunächst protoniert vorliegen, um anschließend das Proton an das Substrat abgeben zu können. Dazu ist erforderlich, dass der pKa-Wert des aktiven Restes geringfügig über 7 liegt. Solche Reste sind oft Aspartat- oder Glutamatreste. Damit sie aktiv sind, muss also ihr pKa-Wert von etwa 4 auf über 7 angehoben sein (Voet & Voet, 2002).

Zusammenfassend ist festzustellen, dass mit den normalen pKa-Werten der meisten titrierbaren Gruppen keine Katalyse möglich ist. Daher sind perturbierte pKa-Werte für die Katalyse häufig essentiell. Um aktive Zentren erfolgreich zu transferieren, ist es also notwendig, sicher zu stellen, dass die katalytisch essentiellen Reste nach der Übertragung geeignete pKa-Werte aufweisen.

2.3.4.1 Berechnung von pKa-Werten

Aus den oben betrachteten Fällen wird ersichtlich, dass es für die Analyse von Reaktionsmechanismen von großem Vorteil ist, die pKa-Werte der aktiven Reste zu kennen. Diese lassen sich experimentell bestimmen oder berechnen, sofern die Struktur des Enzyms bekannt ist.

Die pKa-Werte eines Proteins werden üblicherweise aus berechneten Titrationsreihen abgeleitet (Fogolari et al., 2002). Dazu werden die Protonierungswahrscheinlichkeiten der titrierbaren Gruppen für eine Serie von pH-Werten bestimmt. Mit Hilfe dieser Wahrscheinlichkeiten wird für jede titrierbare Gruppe eine Titrationskurve approximiert. Aus den Kurven lassen sich die pH-Werte ermitteln, bei denen die Protonierungswahrscheinlichkeit bei 50% liegt. Diese pH-Werte sind die gesuchten pKa-Werte.

Es ist aufwändig die Protonierungswahrscheinlichkeiten zu bestimmen. Hierfür müssen zuerst die Interaktionsenergien für alle theoretisch möglichen Protonierungszustände des Proteins berechnet werden. Für n titrierbare Gruppen ergeben sich 2^n Zustände, da jede Gruppe protoniert oder deprotoniert vorliegen kann. Aus den Energien der Zustände lassen sich deren Zustandswahrscheinlichkeiten ableiten. Die Summe der Wahrscheinlichkeiten für alle Zustände, in denen eine titrierbare Gruppe protoniert ist, entspricht dann der Protonierungswahrscheinlichkeit der Gruppe für den betrachteten pH-Wert.

Die Interaktionsenergie für einen Zustand setzt sich zusammen aus den Energien für Desolvatation, Wasserstoffbrücken und Ladungsinteraktion. Hierbei ist vor allem die Berechnung der Ladungsinteraktionen aufwändig. Zum einen muss sie für jeden der 2^n Zustände berechnet werden, zum anderen ist die Berechnung der Interaktionsenergien selbst aufwändig. Der Grund liegt in der unterschiedlich starken Abschirmung der einzelnen Ladungen. Für eine genaue Berechnung muss zusätzlich der Abschirmungsgrad der einzelnen Interaktionen berechnet werden (Bashford & Karplus, 1990).

2.3.4.2 Proteindesign und pKa-Berechnung

Mit EGAD (Pokala & Handel, 2004) lassen sich neben Sequenzoptimierungen auch pKa-Werte berechnen. FDPB_MF (Barth et al., 2007) ist eine weitere Methode, die ähnlich wie EGAD eine Proteindesignmethode für die pKa-Berechnung nutzt. Der

Vorteil dieser Methoden ist, dass die Strukturmodelle, auf denen die pKa-Berechnung beruht, mit flexiblen Seitenketten modelliert werden. Dazu wird das Proteindesign im Seitenkettenoptimierungsmodus (vgl. Kapitel 2.3.1.3) verwendet. So lassen sich Strukturmodelle für das Protein in den verschiedenen Protonierungszuständen berechnen. Aus den Energien für die Modelle wird der pKa-Wert bestimmt (siehe oben). Für ein Protein mit n titrierbaren Gruppen sind dazu 2^n Seitenketten-Optimierungen durchzuführen. Typischerweise liegt der Rechenaufwand dafür im Bereich von Minuten. Durch aufwändige Optimierungsdetails liegt der Berechnungsaufwand bei FDPB_MF für ein einzelnes Protein sogar in der Größenordnung von Tagen.

An dieser Stelle bietet es sich an, den Unterschied zwischen *Vorhersage* und *Optimierung* herauszustellen. Optimierung ist viel aufwändiger, da sie kontinuierliche Vorhersagen für einzelne Optimierungsschritte benötigt. Die beiden genannten Programme benötigen zur *Vorhersage* der pKa-Werte für ein Protein bereits intern mehrere Proteindesignoptimierungen. Bei einer pKa-Wert-*Optimierung* würde dieser Rechenaufwand in jedem Optimierungsschritt entstehen.

2.3.4.3 PKD

PKD (Krieger et al., 2006; Tynan-Connolly & Nielsen, 2006) ist ein Proteindesignprogramm zum Optimieren von pKa-Werten, das in der Arbeitsgruppe von Jens Nielsen entwickelt wurde. Dazu werden für einzelne Reste eines Proteins pKa-Werte angegeben, die durch die Optimierung angenähert werden sollen. Das Programm versucht dann durch gezielte Punktmutationen in der Umgebung den pKa-Wert der zu optimierenden Reste in die gewünschte Richtung zu verändern. Da die pKa-Werte mit einer klassischen Methode berechnet werden, ist die Bewertung einzelner Proteinmodelle relativ aufwändig. Daher wird mit dem Programm nur ein kleiner Teil des Sequenzraums abgetastet, indem nur Einzelmutationen betrachtet werden, die energetisch mit der Struktur verträglich sind. Anschließend werden die Mutationen, die den gewünschten Optimierungseffekt zeigen, miteinander kombiniert. Um Rechenaufwand für diese Mehrfachmutanten zu sparen, werden sie nicht mehr explizit als Kombination modelliert. Stattdessen werden die pKa-Werte aus vorberechneten Daten der Einzelmutanten approximiert. Trotz dieser Vereinfachung weisen die gefunden Lösungen im Vergleich zu vollständigen Berechnung kaum Unterschiede auf, wenn für die geforderten pKa-Werte nur ein kleiner Teil des Proteins mutiert werden muss.

2.3.4.4 PROPKA

Eine interessante Alternative zum Ansatz, pKa-Werte über Titrationskurven zu berechnen, bietet das Programm PROPKA (Li et al., 2005). Entwickelt wurde es in der Arbeitsgruppe von Jan Jensen. Bei PROPKA werden die Perturbationseffekte mit Hilfe

einer semiempirischen Funktion berechnet, die auf einem einfachen physikalischen Modell basiert. Die Funktion wurde so parametrisiert, dass sie für einen Testdatensatz von Proteinen die experimentellen pKa-Werte bestmöglich reproduzieren kann.

Ein Teil der Effekte, die den pKa-Wert einer titrierbaren Gruppe beeinflussen, wird als statischer Hintergrund zusammengefasst. Zum statischen Hintergrund zählen alle Effekte, die unabhängig von den Protonierungszuständen der anderen Gruppen wirken. Dazu zählen der Desolvatationseffekt und Wasserstoffbrücken mit nicht-titrierbaren Gruppen. Die aus dem statischen Hintergrund resultierende pKa-Wert-Verschiebung wird zum Referenz-pKa-Wert der titrierbaren Gruppe addiert. So ergibt sich ein vorläufiger pKa-Wert.

Für lösungsmittelzugängliche Ladungen wird angenommen, dass sie keinen Effekt auf die anderen titrierbaren Gruppen haben, weil das Lösungsmittel diese Ladung abschirmt. Alle anderen Effekte werden als dynamische Effekte modelliert. Dies sind die Wirkungen von Landungen sowie Wasserstoffbrücken zwischen titrierbaren Gruppen, die vom Lösungsmittel abgeschirmt sind. Dynamisch bedeutet hier, dass die Effekte der titrierbaren Gruppen gleichzeitig Protonierungszustände beeinflussen und von ihnen abhängen. Ein Beispiel sind ionisierte, titrierbare Gruppen mit gleichnamigen Ladungen. Deren Ladungen bewirken wechselseitig, dass die pKa-Werte titrierbarer Gruppen in Richtung der neutralen Form verschoben werden. Neutrale titrierbare Gruppen wiederum üben keinen Effekt aufeinander aus.

Der Einfluss dieser dynamischen Effekte wird in einem iterativen Verfahren ermittelt. Dieses Verfahren ersetzt in PROPKA den klassischen Ansatz, alle möglichen Protonierungszustände und die damit verbundenen Energien des Proteins explizit zu berechnen, um daraus die Protonierungswahrscheinlichkeit abzuleiten. Hierzu wird in der ersten Runde auf Grundlage der vorläufigen pKa-Werte ein Protonierungsmodell für das ganze Protein berechnet (die pKa-Werte beschreiben ja eine Protonierungswahrscheinlichkeit). Mit den dynamischen pKa-Perturbationen, die sich aus diesem Protonierungsmodell ergeben, werden neue pKa-Werte berechnet, die anstelle der vorläufigen pKa-Werte in der nächsten Runde verwendet werden. In der nächsten Runde werden die Schritte der ersten Runde wiederholt. Auf diese Weise verstärken sich gleichgerichtete Effekte und der Protonierungszustand beginnt zu konvergieren. Dies bedeutet, dass sich der Protonierungszustand auch durch weitere Runden nicht mehr ändert. Das Programm stoppt bei Konvergenz oder spätestens nach Ablauf einer vorgegebenen Rundenanzahl und gibt die zuletzt berechneten pKa-Werte aus. Dieses Verfahren hat Ähnlichkeiten mit der SCMF-Optimierung (Koehl & Delarue, 1994; Lee, 1994).

Aufgrund der konzeptionell einfachen Modellierung der Effekte und der hieraus resultierenden Vereinfachung der Berechnung ist PROPKA im Vergleich zu Program-

men, die auf Titrationsberechnungen basieren, um mehrere Größenordnungen schneller. Für die Berechnung der pKa-Werte eines Proteins benötigt PROPKA etwa einer Sekunde. Trotz des vereinfachten Berechnungskonzepts ist das Programm erstaunlich genau (Davies et al., 2006; Powers & Jensen, 2006). Berechnete pKa-Werte weichen im Mittel nur um etwa eine pH-Einheit von experimentell bestimmten Werten ab.

2.4 Aufgabenstellung

Ziel dieser Arbeit war es, ein Programm zu entwickeln, das die Übertragung aktiver Zentren von einem Enzym auf ein anders modelliert. Das Programm soll dazu genutzt werden, die Evolution von $(\beta\alpha)_8$-Barrel-Enzymen zu untersuchen. Hierzu sollen mit Hilfe des Programms aktive Zentren zwischen Paaren von Ribulosephosphat-bindenden $(\beta\alpha)_8$-Barrel-Enzymen übertragen werden.

Vier Rahmenbedingungen sollten bei der Modellierung des Transfers von aktiven Zentren besondere Beachtung finden: 1) Proteinstabilität, 2) Ligandenbindung, 3) Ähnlichkeit zum aktiven Zentrum des Vorlageproteins und 4) katalytische Reaktivität durch Optimierung von pKa-Werten. Es gibt gegenwärtig kein Programm, das diesen Anforderungen vollständig entspricht. Allerdings existieren Verfahren, die Teile der Anforderungen abdecken. Daher beruht das in dieser Arbeit entwickelte Programm auf einer hybriden Kombination aus bereits etablierten Verfahren und neuen Ansätzen.

Die Wahrung der Proteinstabilität ist eine Rahmenbedingung, auf der die gesamte Modellierung aufbaut. In Kapitel 2.3.1.5 wurden vier Programme vorgestellt: ROSETTA DESIGN (Kuhlman & Baker, 2000), DEZYMER (Hellinga & Richards, 1991), HERO (Gordon et al., 2003), sowie EGAD (Chowdry et al., 2007). Mit diesen Programmen ist Proteindesign im Hinblick auf die Stabilitätsoptimierung möglich. Unterschiede bestehen nur bei den Optimierungsverfahren und in der Komposition der Energiefunktionen. Der Quellcode ist allerdings nur von ROSETTA DESIGN und EGAD frei verfügbar. Daher wurden in dieser Arbeit nur diese beiden Programme berücksichtigt. Die Programme werden nicht nur für die Stabilitätsoptimierung benötigt. Sie stellen gleichzeitig auch die Komponenten für das Poteindesignkonzept, in das die Ansätze zur Wahrung von weiteren Rahmenbedingungen integriert werden.

Bei der Optimierung der Ligandenbindung müssen zwei Aspekte beachtet werden: Die Ligandenpositionierung und die Ligandenbindungsaffinität. In jeder der oben erwähnten Proteindesignmethoden sind eigens entwickelte Verfahren zur Ligandenpositionierung implementiert (Chowdry et al., 2007; Hellinga & Richards, 1991; Lassila et al., 2006; Zanghellini et al., 2006). In der vorliegenden Arbeit sollen die aktiven Zentren zwischen Ribulosephosphat-bindenden $(\beta\alpha)_8$-Barrel übertragen werden. In diesen Enzymen werden ähnliche Substrate in ähnlicher Orientierung über eine Phosphatbindestelle

gebunden. Daher kann sowohl die Position als auch die Bindungspose des Liganden durch einfache Superpositionierung übertragen werden. Damit sind aufwändige Positionierungsalgorithmen von sekundärer Bedeutung.

Dies gilt jedoch nicht für die Affinitätsoptimierung. Auch hier haben alle Proteindesignprogramme eigene Energieterme für die Berechnung der Wechselwirkungsenergien zwischen Protein und Liganden. Für ROSETTA DESIGN wird dazu die ROSETTA LIGAND Energiefunktion (Meiler & Baker, 2006) aktiviert. Die Methode wurde zwar publiziert, ist aber noch nicht Teil der frei verfügbaren ROSETTA Software. Bei EGAD ist eine Energiefunktion für die Ligandenwechselwirkung zwar implementiert, aber noch nicht als funktionsfähig freigegeben. Daher wurde in dieser Arbeit für die Ligandenbindung das Programm DRUGSCORE (Gohlke et al., 2000) verwendet, das für das Liganden-Docking entwickelt worden ist und dort bereits erfolgreich eingesetzt wird (Wang et al., 2003).

Für die dritte Rahmenbedingung, die Ähnlichkeitsoptimierung, sind bisher keine geeigneten Verfahren publiziert, die direkt verwendbar wären. Dies liegt an der hier vorliegenden speziellen Aufgabenstellung. Es sollen aktive Zentren zwischen sehr ähnlichen Proteinen getauscht werden. Die Hoffnung dabei ist, viele Reste übertragen und äquivalent positionieren zu können, um Einblicke in die Evolution der Funktionen von $(\beta\alpha)_8$-Barrel-Enzymen zu bekommen.

Im Gegensatz zu dieser Aufgabenstellung sind in bisher realisierten Algorithmen die Menge der Reste, die getauscht werden müssen, anders definiert (Bolon & Mayo, 2001; Dwyer et al., 2004; Zanghellini et al., 2006). Um aktive Zentren auf strukturell unähnliche Proteingerüste übertragen zu können, werden nur die katalytisch essentiellen Reste berücksichtigt. Die übrigen Reste der Vorlage werden als nur für Bindung relevant eingestuft und ignoriert. Die Ligandenbindung wird im neuen aktiven Zentrum uniform, dass heißt ohne spezielle Vorlage, optimiert. Methoden, welche die Ähnlichkeit direkt in Betracht ziehen, sind zum Beispiel sequenzbasierte Methoden (Russ et al., 2005; Socolich et al., 2005). Da sich diese nicht ohne weiteres in einen strukturbasierten Ansatz integrieren lassen, wurde hier eine neue Methode implementiert, die auf empirischen Potentialen beruht und im weiteren Verlauf Funktionsdefinition genannt wird. Die Funktionsdefinition wird aus einer Menge von Beispielen für das zu modellierende aktive Zentrum abgeleitet. Ähnliche Ansätze werden zum Beispiel auch bei Homologie-Modellierungsansätzen verwendet (Ginalski, 2006; Sali & Overington, 1994).

Die letzte Rahmenbedingung betrifft die Reaktivität der katalytisch essentiellen Reste. Die Reaktivität kann durch die Optimierung der pKa-Werte von essentiellen Resten des neuen aktiven Zentrums beeinflusst werden. Programme, die hier Verwendung finden sollen, müssen präzise Ergebnisse liefern und gleichzeitig sehr schnell sein. Nur wenn

diese Voraussetzungen erfüllt sind, lässt sich diese Rahmenbedingung sinnvoll in die Optimierung integrieren. Methoden, die pKa-Werte durch berechnete Titrationskurven bestimmen, sind sehr rechenintensiv und daher ungeeignet. Zum Optimieren von pKa-Werten gibt es zwar das Programm PKD (Tynan-Connolly & Nielsen, 2006), das durch geschicktes Kombinieren von Einzelmutationseffekten Lösungen in vertretbarer Zeit findet. Das Programm ist aber nicht mit dem klassischen Proteindesignkonzept verträglich, da beim Proteindesign ein wesentlich größerer Sequenzraum abgetastet werden muss. Die schnellste Methode, die pKa-Werte mit ausreichender Genauigkeit berechnen kann, ist PROPKA (Li et al., 2005; Powers & Jensen, 2006). Daher wurde hier diese Methode verwendet.

Trotz der relativ guten Performanz ist die Originalversion von PROPKA für das Proteindesign nicht nutzbar. Es war daher notwendig, die Geschwindigkeit der Methode durch geeignete Optimierungen um mehrere Größenordnungen zu steigern.

Zusammenfassend kann das Ziel dieser Arbeit wie folgt präzisiert werden: Es soll ein neues Programm zur Übertragung von aktiven Zentren entwickelt werden, dessen Leistung auf der Wahrung von vier Rahmenbedingungen bei der Modellierung beruht. Diese Rahmenbedingungen sollen durch die oben ausgewählten Methoden modelliert werden. Das Basismodul des Programms bilden dabei ROSETTA DESIGN oder EGAD. Sie stellen die folgenden Proteindesignkomponenten: Modellierungseinheit, Energiefunktion und Optimierungsverfahren. Die Methoden DRUGSCORE, PROPKA sowie die Funktionsdefinition sollen modulartig in das Programm integriert werden. Dazu müssen sie an die Proteindesignmethodologie angepasst werden. Die Leistungsfähigkeit des gesamten Programms soll durch ein geeignetes Testverfahren untersucht werden. Schließlich sollen mit dem resultierenden Programm Umwandlungsdesigns für Ribulosephosphat-bindende $(\beta\alpha)_8$-Barrel berechnet und analysiert werden.

3 Material und Methoden

In diesem Kapitel werden verwendete Programme und Verfahren aufgelistet und im Detail beschrieben. Zusätzlich werden Datenquellen aufgeführt, die im Rahmen der Evaluierung benötigt wurden und erläutert, wie diese Daten aufbereitet worden sind.

3.1 Berechnungsverfahren

Die Ergebnisse der vorliegenden Arbeit wurden mit verschiedenen Berechnungsverfahren generiert und analysiert. Im Folgenden werden relevante Berechnungsaspekte im Detail erläutert.

3.1.1 RMSD

Der RMSD (*root mean square deviation*) ist ein Standardmaß um die Abweichung der Atompositionen zwischen zwei Strukturmodellen zu beschreiben. Dazu werden zwei Mengen *A, B* von *n* Atompositionen definiert, die sich paarweise einander zuordnen lassen.

$$RMSD(A,B) = \sqrt{\frac{1}{n}\sum_{1}^{n}\left(\left|a_i - b_i\right|\right)^2}$$

Dabei sind a_i und b_i die Koordinatenvektoren der jeweiligen Atompositionen. Der RMSD-Wert wird als mittlere Koordinatenabweichung für die betrachteten Atompositionen interpretiert und wird üblicherweise in Ångström [Å] angegeben.

Um den RMSD-Wert zwischen zwei Strukturen zu berechnen, gibt es verschiedene Berechnungsmodelle. Für zwei Strukturen mit identischer Sequenz kann der RMSD-Wert über alle Atome berechnet werden. Sind die Sequenzen hingegen unterschiedlich wird der RMSD-Wert nur für Atome äquivalenter Sequenzpositionen bestimmt. Üblicherweise wird der RMSD-Wert in diesen Fällen zwischen den C_α-Atomen oder zwischen allen Rückgratatomen (N, C_α, C, O) berechnet.

Aus diesem Grund muss für eine Beurteilung der Ähnlichkeit zwischen zwei Strukturen neben dem RMSD-Wert auch das zugrunde liegende Berechnungsmodell berücksichtigt werden.

3.1.2 BLOSUM-Score

Um die Ähnlichkeit zwischen zwei alignierten Sequenzen anzugeben, wird oft der Sequenzidentitätswert berechnet. Dieser beschreibt den Anteil an übereinstimmenden Positionen. Eine Alternative zu diesem Verfahren ist der BLOSSUM-Score. Im Gegensatz zur Sequenzidentität berücksichtigt der Wert die Ähnlichkeit aller einander zugeordneten Aminosäuren. Für das Alignment zweier Sequenzen $A=a_1, a_2,\ldots a_n$ und $B=b_1, b_2,\ldots b_n$ ergibt sich der BLOSUM-Score wie folgt:

$$BLOSUM\ -\ Score(A,B) = \frac{1}{n}\sum_{i=1}^{n} BLOSUM_{62}(a_i, b_i)$$

Die Scores $BLOSUM_{62}(a_i,b_i)$ werden der $BLOSUM_{62}$-Substitutionsmatrix (Henikoff & Henikoff, 1992) entnommen.

Höhere Werte für den BLOSUM-Score implizieren dabei ähnlichere Sequenzen. Der Wert für den BLOSUM-Score kann für zwei Sequenzen auch negativ sein, da die Substitutionsmatrix für sehr unähnliche Aminosäurepaare negative Werte aufweist. Da in der vorliegenden Arbeit BLOSUM-Scores nur für Sequenzen gleicher Länge berechnet wurden, mussten mit dem BLOSUM-Score keine Sequenzlücken bewertet werden.

3.1.3 Superpositionierung und strukturbasiertes Sequenzalignment

Für den Vergleich zweier Proteinstrukturen ist es häufig erforderlich, diese vorher zu superpositionieren. Der RMSD-Wert zwischen zwei Strukturen lässt sich beispielsweise erst nach Superpositionierung berechnen.

Die Superpositionierung von zwei Proteinstrukturen wurde in der vorliegenden Arbeit immer mit dem Programm TM-ALIGN (Zhang & Skolnick, 2005) durchgeführt. Dieses Programm bestimmt für beide Strukturen eine Rotationsmatrix und einen Translationsvektor. Mit diesen Ausgaben lässt sich eine Transformation der Atompositionen der zweiten Struktur durchführen. Dadurch kommen beide Strukturen optimal zur Deckung (minimaler RMSD).

Zusätzlich liefert das Programm auch ein strukturbasiertes Sequenzalignment für die Sequenzen der beiden Eingabestrukturen. Um strukturbasierte Sequenzalignments zu generieren wurde daher ebenfalls TM-ALIGN verwendet.

Das Programm TM-ALIGN kann maximal zwei Strukturen superpositionieren. Mussten mehr als zwei Strukturen zusammen superpositioniert werden, so wurde der MULTIPROT-Server (Shatsky et al., 2004) verwendet.

3.1.4 MSAs und Konserviertheit

Die Bedeutung einzelner Aminosäuren für ein Protein lässt sich durch ein entsprechendes MSA ermitteln. Im MSA wird die Bedeutung als Konserviertheit der entsprechenden Spalten sichtbar.

Um MSAs zu generieren wurde das Programm MAFFT (Katoh et al., 2002) verwendet. MAFFT ist in der Lage, sehr umfangreiche Datenmengen (>200 Sequenzen) zu verarbeiten.

Es gibt eine ganze Reihe von Verfahren um die Konserviertheit einzelner MSA-Spalten zu berechnen (Valdar, 2002). In der vorliegenden Arbeit wurde die Konserviertheit mit dem Verfahren von (Karlin & Brocchieri, 1996) in Verbindung mit der $BLOSUM_{62}$-Substitutionsmatrix (Henikoff & Henikoff, 1992) verwendet.

Hierzu wird für eine Spalte *A* in einem MSA zu jeder Kombination von Aminosäuren a_i und a_j in den Zeilen $i < j$ ein normierter BLOSUM-Wert (Division durch die Indentitätswerte für a_i und a_j) berechnet und diese im Anschluss gemittelt:

$$Cons(A) = \sum_{j=1}^{n} \sum_{j>i}^{n} \frac{m(a_i, a_j)}{m(a_i, a_i) m(a_j, a_j)} \times \frac{2}{N/(N-1)}$$

Wobei die Werte $m(a_i, a_j)$ aus der $BLOSUM_{62}$-Substitutionsmatrix stammen. Durch die Normierung mit Hilfe des Nenners bewegt sich der Konserviertheitswert *Cons(A)* zwischen 0 und 1.

3.1.5 CORE-Wert

Die Qualität eines MSAs hängt nicht nur von dem verwendeten Programm, sondern auch von der Eingabe, also von den zu alignierenden Sequenzen ab. Eine Möglichkeit die Qualität von MSAs zu bewerten ist in (Notredame, 2003) beschrieben. Hier werden für die einzelnen MSA-Positionen so genannte CORE-Werte mit dem Programm T-COFFE (Notredame et al., 2000) berechnet.

Für die Berechnung der CORE-Werte werden die Sequenzen eines MSAs zusätzlich paarweise aligniert (PSA). Anschließend wird die Übereinstimmung zwischen MSA und den einzelnen PSAs berechnet. Dieser Übereinstimmung wird für jede Position im MSA mit Werten zwischen 0 und 9 angegeben. Für die Bestimmung der Alignmentqualität für zwei Sequenzen wird in der Vorliegenden Arbeit der CORE-Wert berechnet indem die CORE-Werte der einzelnen Sequenzpositionen gemittelt werden.

3.1.6 Wahrscheinlichkeitsdichten

Im Rahmen der vorliegenden Arbeit war es notwendig, aus Stichproben von Merkmalsausprägungen Wahrscheinlichkeitsdichten abzuleiten, welche die zu Grunde liegende Merkmalsverteilung approximieren. Kann für eine Stichprobe von eindimensionalen Messwerten angenommen werden, dass sie normalverteilt sind, so lässt sich die Verteilung mit einer Gaussfunktion *f* als Wahrscheinlichkeitsdichte angeben:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right)$$

Die dafür benötigten Parameter der Gaussfunktion - Mittelwert μ und Varianz σ - werden dabei aus der Stichprobe ermittelt.

Für die vorliegende Arbeit mussten Wahrscheinlichkeitsdichten für die räumliche Verteilung von Atomen geschätzt werden. Dazu ließ sich das beschriebene Konzept auf einen allgemeinen mehrdimensionalen Fall ausdehnen. Wahrscheinlichkeitsdichen für die Verteilung *n*-dimensionaler Merkmalsvektoren können mit einer multivariaten Gaussfunktion *g* geschätzt werden.

$$g(\bar{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\Sigma\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\bar{x}-\bar{\mu})^T \Sigma^{-1} (\bar{x}-\bar{\mu}) \right)$$

Dabei ist $\bar{\mu}$ der Mittelwertsvektor und Σ die Kovarianzenmatrix, die sich aus der Stichprobe ermitteln lassen. Um die multivariate Gaussfunktionen zu berechnen wurde in der vorliegenden Arbeit die BLOG-Bibliothek verwendet (Milch, 2004).

3.1.7 Zuordnung mit Ungarischer Methode

Wenn für ein modelliertes aktives Zentrum bestimmt werden soll in welchem Grad es der Vorlage entspricht, besteht ein Zuordnungsproblem. Es müssen Strukturdetails des Modells denen der Vorlage zugeordnet werden. Dabei soll die Zuordnung auf eine Weise erfolgen, dass sich die Strukturdetails in der Summe maximal gut entsprechen.

Dieses Zuordnungsproblem lässt sich wie folgt beschreiben: Es sollen *n* Elemente einer Menge *m* Elementen einer anderen Menge paarweise zugeordnet werden, wobei $n<m$ gilt. Jede mögliche Kombination von einander zugeordneten Elementen verursacht dabei Kosten, die es zu minimieren gilt. Die Kosten sind durch eine *n* x *m* Kostenmatrix gegeben.

Die Ungarische Methode (Kuhn, 1955) ist ein Algorithmus, der dieses Zuweisungsproblem in polynomialer Zeit $O(n^3)$ lösen kann. Sie wird in der vorliegenden Arbeit für die Zuordnung verwendet und daher nun genauer beschrieben:

Sollen Elemente aus unterschiedlich großen Mengen einander zugeordnet werden, wird die kleinere Menge um *m-n* „*Dummy*"-Elemente aufgefüllt. Die Kostenmatrix wird für diese Elemente um Kosten für fehlende Zuordnung einzelner Elemente erweitert, so dass in jedem Fall eine quadratische Matrix entsteht.

Der Algorithmus versucht die Kostenmatrix so zu transformieren, dass sie *n* unabhängige Nullen enthält. Unabhängige Nullen, sind Nullen die in der Matrix weder die Spalte noch die Zeile mit weiteren unabhängigen Nullen teilen. Sind *n* unabhängige Nullen vorhanden, ist durch sie eine optimale Zuordnung gegeben, die in Bezug auf die Ausgangskostenmatrix minimal ist.

Der Algorithmus gliedert sich in fünf Schritte:

1. Um Nullen zu generieren wird innerhalb jeder Spalte das Minimum bestimmt und von allen Elementen der Spalte abgezogen. Anschließend wird das Minimum jeder Zeile bestimmt und entsprechend von allen Elementen der Zeile abgezogen.
2. Enthält die so transformierte neue Matrix *n* unabhängige Nullen, ist die optimale Zuordnung bestimmt und der Algorithmus beendet.
3. Ansonsten wird eine minimale Menge *X* von Spalten und Zeilen der aktuellen Kostenmatrix bestimmt, die in der Kombination alle Nullen erfassen.
4. Unter den verbliebenen Elementen wird das Minimum bestimmt. Verbliebene Elemente sind die Elemente der aktuellen Kostenmatrix, welche weder in Spalten noch in Zeilen von *X* vorkommen. Dieser Zahlenwert wird von allen verbliebenen Elementen abgezogen. Außerdem wird er auf alle Elemente addiert, die in X sowohl in einer Spalte als auch in einer Zeile vorkommen.
5. Der Algorithmus wird in Schritt 2 fortgesetzt.

In der vorliegenden Arbeit wurde eine JAVA-Implementation des Algorithmus von (Nedas, 2005) verwendet.

3.1.8 Simulated Annealing Protokoll

In ROSETTA DESIGN (Kuhlman & Baker, 2000) wird für die Optimierung mit SA ein Protokoll verwendet, das für die vorliegende Arbeit übernommen worden ist. Das Protokoll verläuft wie folgt:

Es wird mit einer zufälligen Konfiguration von Rotameren und einer virtuellen Temperatur von 100 gestartet. Die Temperatur wird in zwanzig Schritten auf 0 abgesenkt. Für jede der schrittweisen Temperaturabsenkungen werden so viele Optimierungsschritte ausgeführt, wie insgesamt Rotamere zur Auswahl stehen. Wird innerhalb von vier Temperaturabsenkungen keine verbesserte Lösung gefunden, wird die Temperatur wieder auf 100 angehoben. Auf diese Weise versucht das Protokoll lokale Minima zu vermeiden. Abschließend durchläuft das Protokoll eine Quenchphase, bei der in zufälliger Reihenfolge noch einmal jedes mögliche Rotamer einzeln gesetzt wird, um zu prüfen, ob dessen Auswahl zu einer energetisch niedrigeren Konfiguration führt. Ausgegeben wird schließlich die niederenergetischste Konfiguration, die während des Protokolls gefunden wurde.

3.2 Verwendete Software und Hardware

Im Folgenden wird die Soft- und Hardwareumgebung beschrieben, welche für die vorliegende Arbeit verwendet worden sind.

3.2.1 ROSETTA DESIGN

Das Programm ROSETTA DESIGN (Kuhlman & Baker, 2000) verwendet Rotamere um verschiedene Seitenkettenkonformationen zu generieren. Um die Auflösung zu erhöhen, kann mit dem Programm die Menge der verwendeten Rotamere erweitert werden. Für einzelne Rotamer werden dann zusätzliche Rotamere generiert, die in den ersten beiden Bindungswinkeln um eine Standardabweichung (~10°) variieren (vgl. Kapitel 2.3.1.1). ROSETTA DESIGN wurde im Rahmen dieser Arbeit immer mit dieser Zusatzoption verwendet.

3.2.2 PROPKA

Die Berechnung von pKa-Werten wurde in dieser Arbeit mit dem Programm PROPKA durchgeführt. Das Programm verwendet zur Berechnung der pKA-Werte eine ganze Reihe von Parametern. Dazu zählen Referenz-pKa-Werte der titrierbaren Gruppen, die Parametrisierung des globalen und lokalen Desolvatationsmodells, Parameter für die Wasserstoffbrückenberechnung sowie Parameter für die Ladungsinteraktion. Diese Parameter und die Berechnungsmodelle, für die sie benötigt werden, sind in (Li et al., 2005) im Detail beschrieben.

Für die Entwicklung einer rotamerbasierten Version des Programms wurden sämtliche Parameter und Berechnungsmodelle übernommen (vgl. Kapitel 4.1.4.1).

3.2.3 MODELLER

Für Homologiemodellierungen wurde das Programm MODELLER (Sali & Blundell, 1993) in Version 8.2 verwendet. Um mit dem Programm eine große Anzahl verschiedener Homologiemodellierungen automatisch durchführen zu können, wurde die Modellierungsroutine um eine Schleife ergänzt. Somit das Programm in der Lage für mehrere Eingabesequenzen sukzessiv Modelle zu generieren.

MODELLER kann bei der Strukturmodellierung auch Liganden zu einem gewissen Grad berücksichtigen. Detailierte Wechselwirkungen lassen sich zwar nicht optimieren, der Ligand kann aber als so genannten BLK-Rest im aktiven Zentrum positioniert werden. Auf diese Weise wird die sterische Hinderung der Ligandenatome berücksichtigt. Damit sich diese Funktion im MODELLER nutzen ließ, mussten sowohl die Eingabedaten entsprechend aufbereitet werden als auch die Modellierungsroutine angepasst werden.

3.2.4 Programmierbibliotheken

Proteinstrukturen wurden mit der Programmierbibliothek MBT (*Molecular Biology Toolkit*) geladen und weiterverarbeitet (Moreland et al., 2005). Die Programmierbibliothek bietet umfassende Funktionen zum Umgang mit Proteinstrukturen und erlaubt hierarchischen Zugriff auf die verschiedenen strukturellen Granularitätsebenenen von Proteinkomplexen bis zu einzelnen Atomen.

Sequenzbezogene Auswertungen wurden zum großen Teilen mit der BIO JAVA Bibliothek (Mangalam, 2002) programmiert. Mit dieser Programmierbibliothek ist es möglich Alignments in verschiedenen Formaten einzulesen und weiter zu verarbeiten.

Die BALL-Bibliothek (Kohlbacher & Lenhof, 2000) wurde verwendet, um Seitenketten von Aminosäuren in Proteinstrukturen zu modifizieren. Neue Seitenkettenkonformationen werden dazu mit einer Rotamerbibliothek generiert. Da BALL in der verwendeten Version 1.0 keine rückgratspezifische Rotamerbibliothek unterstützt, wurden die Routinen um eine solche Unterstützung erweitert. Ab Version 1.2 unterstützt BALL rückgratspezifische Rotamerbibliotheken.

3.2.5 Programmiersprachen und Computerausstattung

TRANSCENT wurde in der Programmiersprache JAVA 1.4 mit der Entwicklungsumgebung JBUILDER 2005 (Borland) entwickelt. Anpassungen der verwendeten Hilfsprogramme wurden in der jeweiligen Programmiersprache vorgenommen. (C++ für ROSETTA DESIGN und EGAD und PYTHON für MODELLER)

Die meisten Berechnungen wurden auf einem Linux-Cluster aus 20 *Dell PowerEdge 2550* Computern mit jeweils 2GB Arbeitsspeicher und zwei PentiumIII-Prozessoren (1GHz) durchgeführt. Umwandlungsmodellierungen wurden als ein-Prozessor-Prozesse mit dem Queuingsystem SUN GRID ENGINE (SGE) auf dem Cluster verteilt. Einzelne Umwandlungen benötigten etwa 0.5 bis 1.5GB Speicher und 5 bis 20min Rechenzeit in Abhängigkeit von der Größe des zu optimierenden aktiven Zentrums.

3.2.6 Abbildungen

Die Abbildungen der Proteinstrukturen wurden mit den Moleküldarstellungsprogrammen DEEP VIEW (Guex & Peitsch, 1997) und PYMOL (DeLano, 2002) generiert. Alle anderen Abbildungen, sowie das vorliegende Dokument sind mit OFFICE 2003 (Microsoft) erstellt worden.

3.2.7 Lizenzen

Die meisten im Rahmen der vorliegenden Arbeit verwendeten Programme sind für akademische Verwendung frei verfügbar. Die Programme ROSETTA-DESIGN, MODELLER und MOLOC erfordern zusätzlich eine kostenfreie Registrierung.

Die rotamerbasierte Version des Programms DRUGSCORE beruht auf Potentialen, die aus der CSD-Datenbank (Allen, 2002) abgeleitet wurden. Daher wird eine kostenpflichtige CSD-Datenbank-Lizenz für die Verwendung des Programms benötigt.

3.3 Daten und deren Aufbereitung

Im Rahmen der vorliegenden Arbeit wurden Struktur- und Sequenzdaten für verschiedene Proteine verwendet. Im Folgenden wird beschrieben, wie diese Daten beschafft und aufbereitet worden sind.

3.3.1 Homologe Sequenzen

Für die Homologiemodellung wurden homologe Sequenzen zu einzelnen Proteinen benötigt. Diese wurden aus der PFAM-Datenbank (Sonnhammer et al., 1997) geladen, soweit für die jeweiligen Proteine PFAM-Einträge hinterlegt waren. Um die Suche nach PFAM-Einträgen zu automatisieren, wurde ein Datenbankskript von Matthias Zwick verwendet, das es erlaubt, mit den PDB-Codes einzelner Proteinstrukturen die zugehörigen PFAM-Einträge zu finden. Für den Fall, dass ein PDB-Eintrag einen Proteinkomplex beschrieb, wurde die Ausgabe entsprechend gefiltert.

3.3.2 Aufbereitung der PDB-Dateien

Für die Evaluierung von TANSCENT wurden Proteinstrukturen benötigt. Entweder dienten die Proteine als Vorlage oder als Proteingerüst für den Transfer aktiver Zentren. Zusätzlich wurden Proteinstrukturen als Vorlage für Homologiemodellierungen benötigt.

Um für ein Protein die am besten geeignete Proteinstruktur zu bestimmen, wurde die SCOP-Datenbank (Murzin et al., 1995) verwendet. Die Strukturen für die Testdatensätze wurden mit Hilfe des CULLING-Server (Wang & Dunbrack, 2003; Wang & Dunbrack, 2005) zusammengestellt. Dieser erlaubt diejenige Teilmenge der PDB-Datenbank zu bestimmen, welche eine Reihe von vorher festgelegten Gütekriterien erfüllt.

Das PDB-Datenformat (Bernstein et al., 1977) erlaubt es Strukturen unvollständig, mehrfach oder gar mehrdeutig zu beschreiben. Damit einzelne Strukturen in den verwendeten Testdatensätzen automatisch verarbeitet werden konnten, war es erforderlich, alle verwendeten Strukturen mit folgenden Schritten aufbereiten:

1. Es wurden nur Kristallstrukturen verwendet.
2. Wenn in der PDB-Datei das relevante Protein mehrfach in verschiedenen Aminosäureketten vorlag, entweder als Teil der asymmetrischen Einheitszelle oder als Homodimer, wurde nur die erste Kette verwendet. Lag das relevante Protein als Hetereodimer vor, wurden entsprechend alle anderen Ketten ignoriert.
3. Waren neben dem Liganden von Interesse noch weitere Liganden bzw. Wassermoleküle vorhanden, wurden diese aus der Struktur gelöscht.
4. Wenn für einzelne Atome Alternativpositionen vorlagen (*Occupancy*-Eintrag), wurde jeweils der Eintrag verwendet, für den die höchste Wahrscheinlichkeit angegeben war.
5. Waren für die Strukturen Wasserstoffpositionen angegeben, wurden diese ignoriert.
6. Wenn ein carboxyterminales Sauerstoff-Atom (OXT) fehlte, wurde es entsprechend ergänzt.
7. Waren Strukturen unvollständig, wurden sie, wenn möglich ignoriert, sonst wurden die fehlenden Atome mit dem Programm MODELLER ergänzt (vgl. Kapitel 3.3.3).

Durch diese Schritte wurde eine Menge von aufbereiteten Proteinstrukturen generiert, die sich für die automatisierte Weiterverarbeitung eignete.

3.3.3 Strukturen für die Umwandlungsmodellierungen

Für Umwandlungsmodellierungen werden vollständige Proteinstrukturen mit einem gebundenem Substrat oder einem geeigneten Analogon benötigt.

Die fünf verwendeten Strukturen der Ribulosephosphat-bindenden $(\beta\alpha)_8$-Barrel-Enzyme HisA, HisF, TrpA, TrpC sowie TrpA waren in Bezug auf diese Kriterien teilweise unvollständig.

Zusammen mit Dr. Marco Bocola wurden die fünf Strukturen daher um fehlende Details ergänzt. Für TrpA (1qoq) und TrpC (1lbf) und TrpF (1lbm) war eine Struktur mit Ligand in der PDB-Datenbank abgelegt. Im Falle von HisA (1qo2) und HisF (1thf) wurde der Ligand aus der Struktur des homologen HisF-Enzyms der Hefe (1ox5) durch Superpositionierung übernommen.

Fehlende Schleifenbereiche der Ausgangsstrukturen wurden mit Hilfe des Programms MODELLER durch eine vergleichende Modellierung mit homologen Strukturen vervollständigt. Die so erzeugten vollständigen Referenzstrukturen mit gebundenem Substrat wurden mit dem Programm MOLOC (Gerber & Müller, 1995) energieoptimiert, um die ideale Platzierung der Seitenketten um das Substrat in der Bindetasche zu gewährleisten. Für die Optimierung von Wasserstoffbrücken ohne zusätzliche Desolvatationskorrektur wurde für das Kraftfeld ein in dieser Situation üblicher Korrekturfaktor von 1.786 für H-Brücken verwendet (Seitz et al., 2007).

4 Ergebnisse

Im Rahmen dieser Arbeit wurde das Programm TRANSCENT entwickelt, das automatisiert eine optimale Übertragung von aktiven Zentren modellieren kann. Die Module, auf denen das Programm beruht, wurden mit einem Gewichtsfindungsprozess aufeinander abgestimmt. Das Programm arbeitet auf der Basis von Homologiemodellen, daher wurde ein Modellierungskriterium bestimmt, mit dem die Qualität der verwendeten Modelle beurteilt werden kann. Zur Evaluierung wurden mit dem Programm die Übertragung von aktiven Zentren zwischen Ribulosephosphat-bindenden $(\beta\alpha)_8$-Barrel-Enzymen modelliert.

In diesem Kapitel werden der Aufbau und die Arbeitsweise des Programms erläutert und das Modellierungskriterium vorgestellt. Anschließend werden die Ergebnisse des Gewichtsfindungsprozesses und der Evaluierung präsentiert.

4.1 Das Programm TRANSCENT

TRANSCENT setzt sich aus vier Modulen zusammen. Jedes Modul stellt die Einhaltung einer anderen Rahmenbedingung bei der Modellierung eines neuen aktiven Zentrums sicher. Das erste Modul sorgt für die Proteinstabilität, das zweite Modul für optimale Ligandenbindung, das dritte Modul für die Ähnlichkeit des Modells zur Vorlage und das vierte Modul für optimale pKa-Werte. Das Programm arbeitet als Proteindesignprozess. Hierbei wird die Besetzung der Positionen des neuen aktiven Zentrums mit Aminosäuren und die Ausrichtung ihrer Seitenketten optimiert. Die einzelnen Module beeinflussen dabei diesen Optimierungsprozess gleichzeitig.

4.1.1 Modul 1: Proteinstabilität

Hinreichende Stabilität des resultierenden Proteins ist die erste Voraussetzung, um aktive Zentren zu übertragen. Die Austausche, die in ein Strukturgerüst eingeführt werden, müssen energetisch verträglich sein.

Es kann davon ausgegangen werden, dass ein Proteinmodell, das mit einer hohen Energie bewertet wird, kein realistisches Modell ist. Hohe Energien werden in realen Proteinen durch Bewegungen der Proteinstruktur kompensiert. Es ist daher notwendig, beim Modellieren eines neuen aktiven Zentrums die notwendigen Austausche unter Erhalt der Stabilität einzuführen. Nur wenn sich die neuen Reste energetisch verträglich integrieren lassen, kann angenommen werden, dass das resultierende Modell die Struktur des Proteins geeignet repräsentiert.

Für TRANSCENT wird Proteindesign zur Modellierung der Proteinstabilität verwendet. Dazu werden die drei Grundkomponenten des Proteindesigns benötigt. Diese sind die Modellierungseinheit, die Energiefunktion und ein Optimierungsverfahren. Das Proteindesign-Modul bildet zusätzlich den Rahmen für die Programmstruktur. Die anderen Module werden sukzessiv in das Proteindesignkonzept integriert.

4.1.1.1 Zwei Optionen für die Stabilisierung: ROSETTA DESIGN und EGAD

In das Stabilitätsmodul wurden zwei Programme integriert, die von verschiedenen Arbeitsgruppen entwickelt wurden. Diese Programme sind ROSETTA DESIGN und EGAD. ROSETTA DESIGN (Kuhlman & Baker, 2000) ist als monolithisches Anwendungsprogramm konzipiert, das mit einer Aufgabenstellung als Eingabe gestartet wird und das anschließend eine Lösung ausgibt. Ein direkter Zugriff auf das interne Datenmodell, also auf die interne Rotamertabelle und die Energietabelle, ist weder per Programmierschnittstelle, noch über eine Datenexportfunktion vorgesehen. Da es aber für die geplante Integration des Programms in TRANSCENT notwendig war, auf das Datenmodell zuzugreifen, hat Nils Enkler im Rahmen seiner Diplomarbeit eine Exportfunktion für das Datenmodell in das Programm integriert (Enkler, 2006).

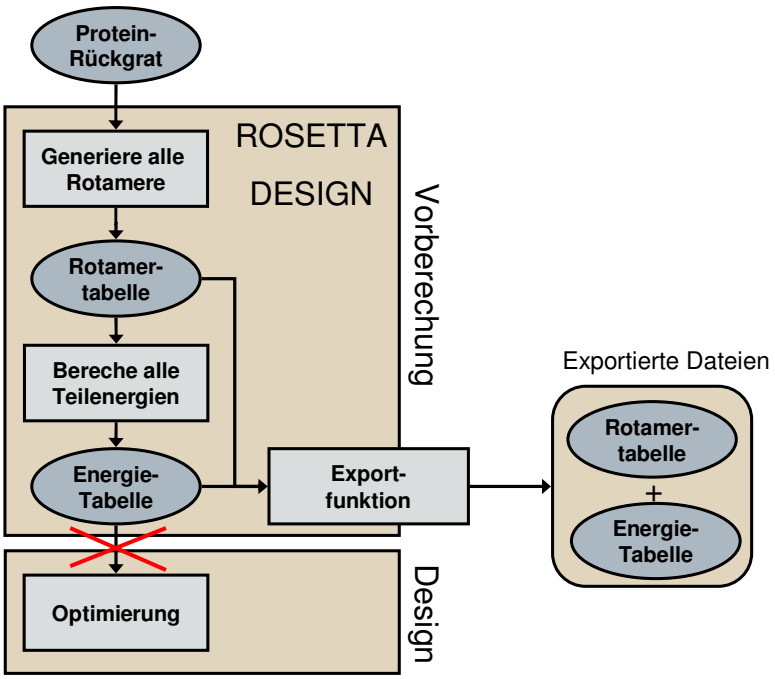


Abbildung 10: Export der Energie- und Rotamertabelle von ROSETTA DESIGN
Das Datenmodell von ROSETTA DESIGN besteht aus einer Rotamertabelle und einer Energietabelle. Um auf diese Daten von außen zuzugreifen, wurde das Programm um eine Exportfunktion erweitert (Enkler, 2006), die es erlaubt, das Datenmodell in Dateien zu exportieren. In diesem Fall wird die Optimierung des Designproblems nicht angestoßen.

Verwendet wird diese Schnittstelle nach folgendem Schema: ROSETTA DESIGN wird mit den Eingaben für ein konkretes Proteindesignproblem gestartet. Im ersten Schritt generiert die Modellierungseinheit von ROSETTA DESIGN die Rotamertabelle und daraus mit Hilfe der Energiefunktion die Energietabelle. Mit Hilfe der Exportfunktion werden die Einträge der Rotamer- und Energie-Tabelle in Dateien gespeichert und auf diese Weise anderen Programmen zur Verfügung gestellt (vgl. Abbildung 10).

Das zweite Proteindesignprogramm, das alternativ im Modul für die Stabilität verwendet wird, ist EGAD (Pokala & Handel, 2004; Pokala & Handel, 2005). Zu diesem Programm gibt es eine Programmierbibliothek, EGAD LIBRARY (Chowdry et al., 2007). Alle für das Proteindesign benötigten Komponenten werden in EGAD LIBRARY als „Baukasten“ zur Verfügung stellt. Zusätzlich sind bereits Schnittstellen implementiert, die es erlauben Rotamertabellen und Energietabellen als Dateien zu exportieren. Daher war es einfach, eine Routine zu entwickeln, die für ein übergebenes Proteinrückgrat die Rotamer- und die Energietabelle berechnet und exportiert.

Damit stehen wahlweise das Datenmodell von ROSETTA DESIGN oder EGAD in Form von Dateien für TRANSCENT zur Verfügung.

4.1.1.2 Modellierungseinheit, Energiefunktion und Optimierungsverfahren

TRANSCENT arbeitet mit ROSETTA DESIGN und EGAD in folgender Weise zusammen: Die Proteinstruktur, auf der ein neues aktives Zentrum etabliert werden soll, wird von TRANSCENT wahlweise an eines der beiden Programme weitergegeben. Dieses berechnet die Tabellen und exportiert sie als Dateien. Mit Hilfe spezifischer Importfunktionen werden die Datensätze eingelesen und in TRANSCENT in ein einheitliches Datenmodell übertragen. Daher können die Daten unabhängig von der Herkunft weiter verwendet werden (vgl. Abbildung 11).

Für ein vollständiges Proteindesignframework benötigt TRANSCENT ein Optimierungsverfahren. Dazu kommen hier vor allem zwei Ansätze in Frage, nämlich DEE und SA. DEE kann im betrachteten Fall als Optimierungsverfahren nicht verwendet werden, da die Module für die Ähnlichkeits- und pKa-Wert-Optimierung die Energiefunktion um Multikörper-Terme (vgl. Kapitel 4.1.3 und 4.1.4) erweitern. Im Gegensatz dazu stellt SA keine einschränkenden Anforderungen an die Energiefunktion und ist auch mit einer Mehrkörper-Energiefunktion vereinbar, da während der Optimierung immer die Energien einer vollständigen Sequenz von Rotameren betrachtet werden. Daher wurde SA als Optimierungsverfahren gewählt.

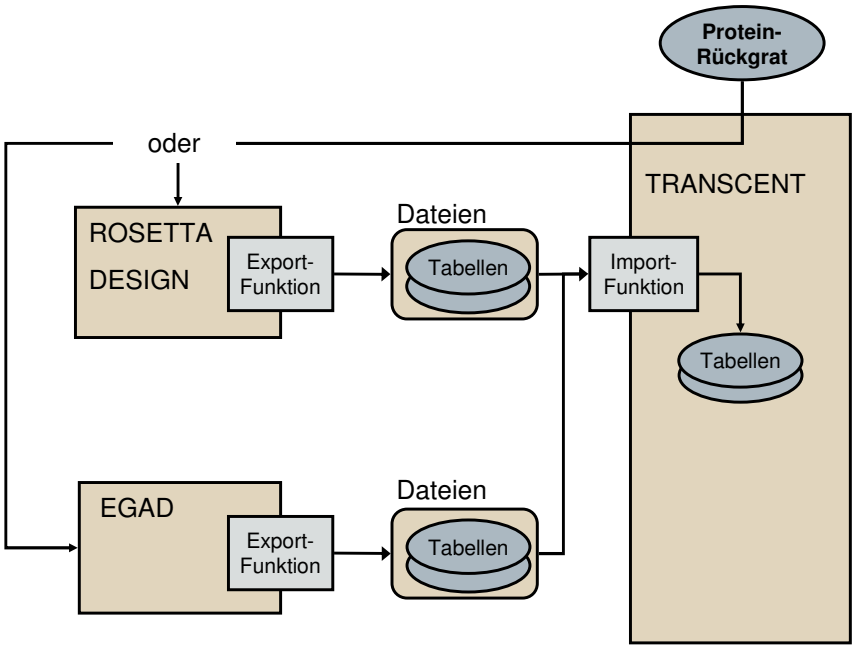


Abbildung 11: Uniforme Verwendung der EGAD und ROSETTA DESIGN Datenmodelle
TRANSCENT stößt das Berechnen der Rotamer- und Energietabelle entweder in ROSETTA DESIGN oder in EGAD an und importiert diese dann in Form von Dateien. Die Herkunft der Daten ist für den weiteren Verlauf des Programms unerheblich, da mittels spezifischer Importfunktionen für eine einheitliche Datenrepräsentation in TRANSCENT gesorgt wird.

Da die importierten Tabellen die Modellierungseinheit und die Energiefunktion des jeweiligen Programms repräsentieren (vgl. Kapitel 2.3.1.4), wird für TRANSCENT keine eigene Variante dieser beiden Komponenten benötigt.

Das SA-Protokoll wurde von ROSETTA DESIGN übernommen, da dieses Protokoll bereits für das Proteindesign optimiert worden ist. Die Optimierungsdauer dieses Protokolls ist von der Problemgröße abhängig. Es werden 100-mal mehr Annealing-schritte berechnet, als Rotamere zur Auswahl stehen. Für ein Proteindesignproblem mit 10.000 Rotameren ergeben sich so eine Millionen Schritte.

Wie geschildert, bildet das erste Modul ein vollständiges Proteindesignprogramm, das aus Modellierungseinheit, Energiefunktion und Optimierungsverfahren besteht (vgl. Kapitel 2.3.1). Die Modellierungseinheit wird durch eine importierte Rotamertabelle repräsentiert, die Energiefunktion durch eine importierte Energietabelle. Da als Optimierungsverfahren das SA-Protokoll von ROSETTA DESIGN implementiert wurde, stellt das Modul eine Reimplementation des ROSETTA DESIGN Programms dar, wenn die Tabellen von ROSETTA DESIGN stammen. Wird TRANSCENT mit diesem Modul alleine betrieben, arbeitet es als Proteindesignverfahren, mit dem sich auf Stabilität optimierte Sequenzen für ein gegebenes Proteinrückgrat berechnen lassen.

4.1.2 Modul 2: Ligandenbindung

Bindung ist eine notwendige Voraussetzung für die Katalyse. Nur wenn der Ligand gebunden ist, kann er mit dem Protein wechselwirken. Wird die Katalyse betrachtet, schließt die Bindung die richtige Orientierung des Liganden im aktiven Zentrum mit ein.

Um aktive Zentren zu transferieren, ist es also notwendig, sicherzustellen, dass die Ligandenbindung im Zielprotein möglich ist. Das Modul für die Stabilisierung berücksichtigt die Ligandenbindung nicht, da in diesem Modul nur die Wechselwirkungen der Seitenketten modelliert werden. Es ist sogar wahrscheinlich, dass bei ausschließlicher Optimierung der Proteinstabilität die Ligandenbindung verschlechtert wird. Denn um die Stabilität zu erhöhen, wird das Protein so dicht wie möglich gepackt und es werden Kavitäten soweit möglich geschlossen. Daher würden auch aktive Zentren, die in der Regel eine Kavität darstellen, durch tendenziell große Seitenketten ausgefüllt. In solchen Fällen ist aber die Ligandenbindung oft schon aus sterischen Gründen nicht mehr möglich. TRANSCENT steuert diesem Effekt mit dem zweiten Modul entgegen, das die Ligandenbindung optimiert.

4.1.2.1 Positionierung des Liganden

Bei der Positionierung eines Liganden in einem aktiven Zentrum müssen drei Aspekte berücksichtigt werden. Diese sind: 1) Die Suche nach einer Bindestelle, 2) die Orientierung des Liganden in der gefundenen Bindestelle und 3) die Bindungspose des Liganden. Bei der Bindestellensuche werden Regionen im Protein identifiziert, die als neues aktives Zentrum in Betracht gezogen werden können. Bei der Ligandenorientierung wird festgelegt, wie der Ligand relativ zur Bindestelle auszurichten ist. Die Bindungspose beschreibt die Konformation des positionierten Liganden.

Da mit TRANSCENT aktive Zentren zwischen Proteinen mit gleichem Faltungstyp übertragen werden, ist sowohl die Suche nach einer Bindestelle, als auch die Auswahl der richtigen Konformation des Liganden erheblich vereinfacht. Es wird davon ausgegangen, dass sich das neue aktive Zentrum an gleicher Stelle und in gleicher Orientierung wie in der Vorlage modellieren lässt (vgl. Kapitel 2.3.2).

Als Katalyse-relevante Bindungsposen werden Übergangszustände von Substrat zu Produkt angesehen (Olsson et al., 2006; Schramm, 2005). Diese zu generieren, geht über den Rahmen dieser Arbeit hinaus. Daher wird vorausgesetzt, dass der Ligand, der in der Vorlagestruktur gegeben ist, sich bereits in einer Katalyse-relevanten Konformation befindet. TRANSCENT superpositioniert die Vorlagestruktur mit der Modellstruktur und überträgt anschließend den Liganden von der Vorlage ins Modell. Hierbei wird die Konformation des Liganden beibehalten. Es können aber auch andere Ligandenkonformationen oder Liganden an alternativen Bindestellen übergeben werden.

4.1.2.2 Ligandenbindung als Selbstenergieterm

Wenn der Ligand, dessen Bindung optimiert werden soll, zusammen mit dem Proteinrückgrat als starre Konfiguration von Atomen gegeben ist, so lassen sich die Wechselwirkungen zwischen den Aminosäureseitenketten und dem Liganden konzeptionell in ähnlicher Weise zusammenfassen, wie die Wechselwirkungen zwischen den Seitenketten und dem Rückgrat (vgl. Kapitel 2.3.1.2). Dazu wird für jedes Rotamer die Wechselwirkungsenergie zum Liganden berechnet, analog zur Berechnung der Wechselwirkungsenergie von Rotameren und starrem Rückgrat. Eine solche Energie wird Selbstenergie genannt, da sie für jedes Rotamer separat, also unabhängig von anderen Rotameren berechnet werden kann. Zusätzliche Selbstenergien lassen sich auf sehr einfache Weise in den Proteindesignprozess integrieren. Sie werden zu den bereits gespeicherten Selbstenergie-Werten der Energietabelle hinzuaddiert. Darüber hinaus sind keine Anpassungen erforderlich.

Ein weiterer Aspekt erleichtert die Integration eines Moduls zu Ligandenbindung in den Proteindesignprozess: Optimale Bindung des Liganden wird durch die Minimierung der Bindungsenergie erreicht. Wenn die Bindungsenergie der Energietabelle hinzugefügt worden ist, impliziert die Optimierung der Stabilität auch die Optimierung der Bindung.

4.1.2.3 Bindungsenergieberechnung: DRUGSCORE

Voraussetzung für die Berechnung der Wechselwirkungsenergien zwischen Seitenketten und Liganden ist die Verfügbarkeit einer entsprechenden Energiefunktion. Für TRANSCENT wurde hierfür DRUGSCORE (Gohlke et al., 2000) verwendet. DRUGSCORE ist ein empirisches Potential zu Bewertung von Protein-Ligand-Wechselwirkungen, welches in der Arbeitsgruppe von Gerd Klebe entwickelt wurde. Um es für das Proteindesign verwenden zu können, muss DRUGSCORE aber einzelne Rotamere bewerten können. Hierfür müssen die Wechselwirkungsenergien zwischen dem Liganden und einzelnen Rotameren separat berechnet werden können.

Gerd Neudert aus der Arbeitsgruppe von Gerd Klebe, hat deshalb für TRANSCENT eine angepasste Version von DRUGSCORE entwickelt. Das Programm erwartet die Struktur einen Liganden und eine Rotamertabelle als Eingabe. Es liefert dann die Interaktionsenergie zwischen dem Liganden und einzelnen Rotameren als Tabelle zurück. Die Interaktionsenergien werden von TRANSCENT eingelesen und der Energietabelle für das Proteindesign hinzugefügt. Dazu wird für jedes Rotamer die DRUGSCORE-Energie mit dem Selbstenergiewert verrechnet, der bereits in der Energietabelle steht (vgl. Abbildung 12).

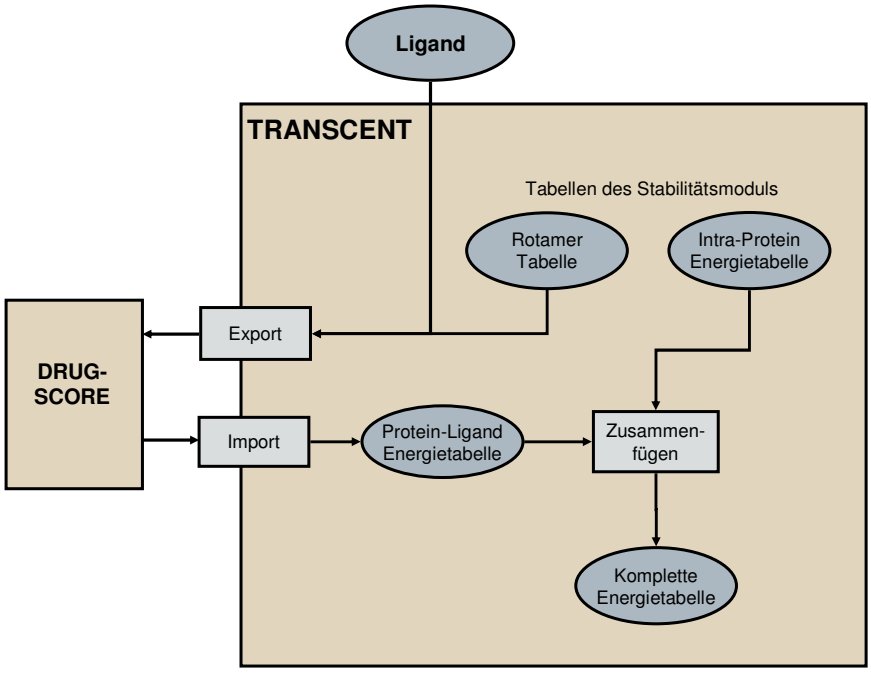


Abbildung 12: Berechnung der Ligandenwechselwirkung mit DRUGSCORE
Das Proteindesign-Modul liefert die Rotamertabelle und die Energietabelle. In der Energietabelle sind noch keine Wechselwirkungen mit dem Liganden gespeichert. Diese werden durch das Programm DRUGSCORE berechnet. Dazu werden dem Programm Ligand und Rotamertabelle übergeben. Die durch DRUGSCORE berechneten Interaktionsenergien werden importiert und der Energietabelle hinzugefügt.

In DRUGSCORE wird auch der Verlust an Oberflächenzugänglichkeit durch die Ligandenbindung beurteilt. Da das Programm Oberflächen aber nur für vollständige Proteine berechnen kann, wird bei der rotamerbasierten Verwendung von DRUGSCORE die Oberflächenberechnung abgeschaltet. Für die Bewertung werden dann nur die abstandsabhängigen Potentiale verwendet.

Mit Hilfe dieses Moduls ist TRANSCENT in der Lage, Proteinstabilität und Ligandenbindung für ein Modell gleichzeitig zu optimieren.

4.1.3 Modul 3: Die Funktionsdefinition

Um ein aktives Zentrum von einem Protein auf ein anderes übertragen zu können, ist es notwendig, in einem ersten Schritt alle Details zu beschreiben, welche für die enzymatische Funktion relevant sind. Üblicherweise umfasst eine derartige Definition die katalytisch essentiellen Reste und deren räumliche Orientierung zum Substrat (vgl. Kapitel 2.3.3). Diese Informationen bilden die Funktionsdefinition. Ein Proteinmodell wird anschließend so optimiert, dass es der Funktionsdefinition möglichst genau entspricht.

Für TRANSCENT wurde das Konzept der Funktionsdefinition erweitert. Da mit dem Programm aktive Zentren zwischen strukturell ähnlichen Proteinen getauscht werden sollen, ist die Wahrscheinlichkeit hoch, dass sich neben den katalytisch essentiellen Resten auch weitere Reste relativ einfach übertragen lassen. Die Funktionsdefinition beschränkt sich daher nicht alleine auf die essentiellen Reste, sondern schließt auch andere Reste ein. Dazu gehören zum Beispiel solche, die für die Bindung relevant sind. Diesem Vorgehen liegt folgende Annahme zu Grunde: Je mehr ein Proteinmodell der Funktionsdefinition entspricht, desto ähnlicher ist es der Vorlage und desto wahrscheinlicher beschreibt es ein aktives Protein mit der gewünschten Funktion.

Die Funktionsdefinition für TRANSCENT wird automatisch erzeugt. Hierfür generiert und analysiert das Programm zuerst eine Strukturbibliothek. Die Strukturbibliothek beinhaltet eine große Menge an Beispielstrukturen für das zu transplantierende aktive Zentrum. Da die Menge der bekannten Strukturen oft nicht ausreicht, wird die Strukturbibliothek durch Homologiemodelle erweitert. Die Beispielstrukturen werden auf funktionsrelevante Merkmale untersucht. Dabei wird angenommen, dass der Konserviertheitsgrad der Merkmale deren funktionale Relevanz beschreibt. Aus den gefundenen Gemeinsamkeiten werden wissensbasierte Potentiale abgeleitet und als Funktionsdefinition zusammengefasst.

Mit Hilfe der Potentiale aus der Funktionsdefinition kann bewertet werden, wie genau ein betrachtetes Modell den Vorlagen entspricht. Daher wird die Funktionsdefinition als Modul in TRANCENT integriert, um die Ähnlichkeit zur Vorlage optimieren zu können. Im Folgenden wird der Aufbau dieses Moduls genauer beschrieben.

4.1.3.1 Funktionsdefinition aus Wasserstoffbrücken-bildenden Gruppen

Die Funktionsdefinition wird aus den Gemeinsamkeiten von aktiven Zentren der Strukturbibliothek abgeleitet (Abbildung 13). Bei katalytisch relevanten Resten ist vor allem die Position und Ausrichtung von Wasserstoffbrücken-bildenden Gruppen (HB-Gruppen) relevant. Diese Gruppen agieren als Donor oder Akzeptor für Wasserstoffbrücken zum Liganden. Neben den katalytischen Resten gibt es auch bindungsvermittelnde Reste mit HB-Gruppen. Daher beschreibt die Funktionsdefinition die Verteilung und Konserviertheit aller HB-Gruppen im aktiven Zentrum.

Durch die Verteilung aller HB-Gruppen wird indirekt auch die Verteilung aller Reste ohne HB-Gruppen, (d.h. der hydrophoben Reste) beschrieben, da sich diese auf den verbleibenden Raum verteilen. Dadurch schließt die Funktionsdefinition implizit auch Informationen über hydrophobe Bereiche ein.

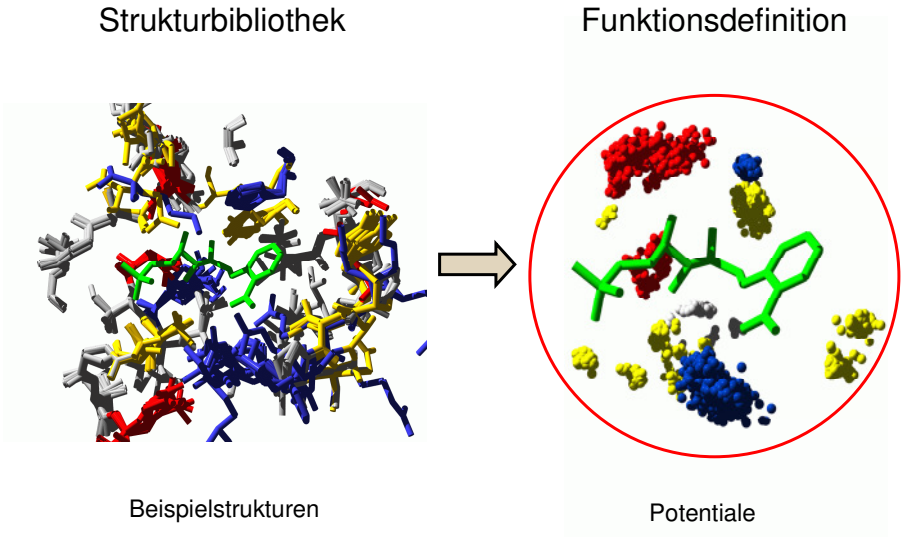


Abbildung 13: Ableitung der Funktionsdefinition aus einer Strukturbibliothek
Um neu modellierte aktive Zentren in Bezug auf Ähnlichkeit zur Vorlage optimieren zu können, werden Potentiale aus einer Strukturbibliothek abgeleitet. Diese bilden dann die Funktionsdefinition. In der Abbildung ist dieser Prozess durch ein Beispiel veranschaulicht. Links sind aktive Zentren von Beispielstrukturen aus der Strukturbibliothek von TrpF überlagert. Das Beispiel ist stark vereinfacht, es werden nur 10 Einträge aus der Strukturbibliothek gezeigt. Tatsächlich besteht die Strukturbibliothek aus über hundert Strukturen. Aus den Beispielstrukturen wird die Verteilung von Wasserstoffbrücken-bildenden Gruppen (HB-Gruppen) abgeleitet. Rechts sind die räumlichen Verteilungen der HB-Gruppen relativ zum Liganden durch Punktwolken angedeutet. Aus diesen Verteilungen werden die Potentiale für die Funktionsdefinition abgeleitet. Die Färbung zeigt Arg, His sowie Lys in blau. Asp und Glu sind rot eingefärbt. Asn, Cys, Gln, Ser, Thr sowie Tyr sind gelb und der Ligand ist grün eingefärbt. Die hydrophoben Reste sind grau dargestellt.

Um eine Funktionsdefinition zu generieren, werden für alle Beispielstrukturen die Reste bestimmt, deren HB-Gruppen eine Wasserstoffbrücke in Richtung des Liganden ausbilden können. HB-Gruppen verschiedener Beispielstrukturen, deren Reste sich an strukturell äquivalenten Positionen befinden, werden als Cluster zusammengefasst. Anschließend wird für jedes Cluster analysiert, aus welchen HB-Gruppen es sich zusammensetzt und wie diese räumlich verteilt sind. Aus diesen Informationen werden Potentiale berechnet.

4.1.3.2 Strukturbibliothek mit Homologiemodellen

Die Potentiale der Funktionsdefinition beschreiben die Variabilität von HB-Gruppen in einem aktiven Zentrum. Diese Variabilität gibt den Spielraum zum Modellieren vor. Um die Variabilität geeignet analysieren zu können, muss eine ausreichend große Menge an Beispielstrukturen verfügbar sein. Beispielstrukturen sind dabei homologe Proteine, d.h.

solche aus verschiedenen Organismen. Häufig enthält die PDB-Datenbank für ein betrachtetes Protein nur eine Struktur. Selten sind es mehr als zehn Strukturen.

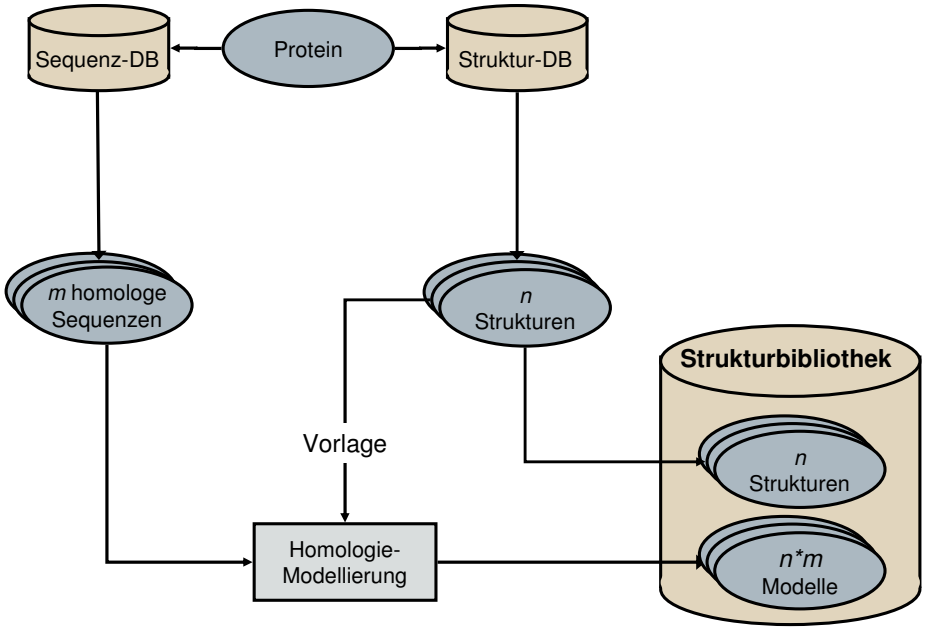


Abbildung 14: Flussdiagramm für die Erzeugung einer Strukturbibliothek
Das Diagramm beschreibt den Prozess, der für ein Protein eine Strukturbibliothek erstellt. Dazu werden bekannte 3D-Strukturen verwendet, aber auch Strukturmodelle, die per Homologiemodellierung erzeugt werden.

Um die Datenbasis zu erweitern, werden in TRANSCENT Strukturmodelle aus homologen Sequenzen erzeugt. Die Strukturen aus der PDB-Datenbank dienen dabei als Vorlage. Die homologen Sequenzen stammen aus der PFAM-Datenbank (Sonnhammer et al., 1997). Modelliert werden die Strukturen mit dem Homologiemodellierungsprogramm MODELLER (Sali & Blundell, 1993). Die generierten Modelle werden mit den Strukturen aus der PDB-Datenbank zusammengeführt und bilden zusammen die Datenbasis der Strukturbibliothek (Abbildung 14).

Als Zuordnungsvorlage dient bei TRANSCENT ein MSA. Durch das MSA ist es möglich, einzelne Positionen der verschiedenen Strukturen einander zuzuordnen. Jede Spalte des MSAs beschreibt zueinander äquivalente Positionen in den Strukturen. Mit „Position in der Strukturbibliothek“ ist in dieser Arbeit immer eine Menge von einander äquivalenten Positionen der einzelnen Strukturen gemeint. Die Reste an einer Position der Strukturbibliothek sind also alle Reste aus den verschiedenen Strukturen, die im MSA in einer Spalte gemeinsam aufgelistet werden.

4.1.3.3 Ableitung der Potentiale

Für die Funktionsdefinition werden aus Verteilungen von HB-Gruppen in der Strukturbibliothek wissensbasierte Potentiale abgeleitet (Sippl, 1990; Sippl, 1995). Mit diesen Potentialen soll beurteilt werden, wie ähnlich das aktive Zentrum eines Proteinmodells zur Menge der Vorlagen ist (vgl. 2.3.3.1).

Die Grundlage für die Potentiale bilden die Strukturen der Strukturbibliothek. Die aktiven Zentren dieser Strukturen dienen als Vorlage für das zu modellierende aktive Zentrum. Dazu wird für die räumliche Anordnung der HB-Gruppen der aktiven Zentren eine Wahrscheinlichkeitsverteilung berechnet ($f_{Vorlage}$). Zusätzlich wird eine erwartete Wahrscheinlichkeitsverteilung für HB-Gruppen-Anordnungen aufgestellt ($f_{Erwartet}$). Mit diesen Wahrscheinlichkeitsverteilungen lässt sich dann das wissensbasierte Potential definieren.

$$E_{Funktionsdef}(x) = -\ln\left(\frac{f_{Vorlage}(x)}{f_{Erwartet}(x)}\right)$$

Dabei bezeichnet x ein aktives Zentrum, für das durch die Funktion beurteilt wird, ob es eher den Vorlagen ($f_{Vorlage}$) oder eher einer allgemeinen Verteilung entspricht ($f_{Erwartet}$).

Die beiden Wahrscheinlichkeitsverteilungen werden aus den relativen Häufigkeiten der HB-Gruppen in zwei verschiedenen Stichproben geschätzt. Die Stichprobe für $f_{Vorlage}$ besteht dabei aus den Strukturen der Strukturbibliothek. Für $f_{Erwartet}$ wird eine Stichprobe von Strukturen benötigt, bei denen die HB-Gruppen im aktiven Zentrum gemäß der allgemeinen Verteilung für HB-Gruppen verteilt sind. Für solche Proteine gibt es aber weder Struktur- noch Sequenzbeispiele. Proteine mit einer unspezifischen HB-Gruppen-Verteilung sind in der Regel inaktiv und würden dem Selektionsdruck erliegen. Daher werden unspezifische HB-Gruppen-Verteilungen mit Hilfe einer Rotamerbibliothek auf die Proteingerüste der Vorlagen modelliert und als Stichprobe für $f_{Erwartet}$ verwendet.

Bei der Schätzung der Verteilungen $f_{Vorlage}$ und $f_{Erwartet}$ werden nur Positionen der Strukturbibliothek betrachtet, bei denen wenigstens eine Seitenkette eine HB-Gruppe besitzt und eine Wasserstoffbrücke zum Liganden bilden kann. Alle HB-Gruppen einer solchen Position werden räumlich als Cluster zusammengefasst (vgl. Abbildung 13).

Im weiteren Vorgehen wird angenommen, dass einzelne Positionen und die damit assoziierten Cluster voneinander unabhängig betrachtet werden dürfen. Diese Approximation wird durch die anderen Module von TRANSCENT gerechtfertigt, die bereits die Abhängigkeiten der Positionen berücksichtigen. Durch diese Unabhängigkeit lässt sich die gesamte Wahrscheinlichkeitsverteilung für alle HB-Gruppen in ein Produkt von lokalen Wahrscheinlichkeitsverteilungen $f_{Vorlage,i}$ und $f_{Erwartet,i}$ für HB-Gruppen an den einzelnen Positionen i zerlegen. Mit den lokalen Wahrscheinlichkeitsverteilungen lassen

sich dann Teilpotentiale definieren, die separate Regionen des aktiven Zentrums beschreiben. Das Gesamtpotential lässt sich auf Grund der Logarithmierungen der Wahrscheinlichkeitsverteilungen als Summe von *n* Teilpotentialen formulieren:

$$E_{Funktionsdef}(x) = -\ln\left(\frac{\prod_{i=1}^{n} f_{Vorlage,i}(x)}{\prod_{i=1}^{n} f_{Erwartet,i}(x)}\right) = \sum_{i=1}^{n} -\ln\left(\frac{f_{Vorlage,i}(x)}{f_{Erwartet,i}(x)}\right)$$

Die HB-Gruppen einer Position sind nicht immer über alle Vorlagestrukturen konserviert. Daher muss die Wahrscheinlichkeitsverteilung auch den Fall, dass in einem Modell eine HB-Gruppe nicht vorhanden ist, differenziert bewerten. Die Bewertung hängt dann davon ab, wie viele verschiedene Aminosäuretypen an dieser Position eine HB-Gruppe zum Liganden ausbilden bzw. wie häufig an der Position keine HB-Gruppe vorhanden ist. Für eine Positionen *i* müssen daher Potentiale für zwei Fälle separat ermittelt werden: das Potential der Apolarität und das Potential der HB-Gruppen. Mit den Potentialen lässt sich dann sowohl ein Modell mit einer HB-Gruppe in der betrachteten Region bewerten, als auch ein Modell, welches in der Region keine HB-Gruppe aufweist.

Um die Apolarität einer Position *i* in der Strukturbibliothek zu beschreiben, wird die relative Häufigkeit der Reste bestimmt, die keine HB-Gruppe zum Liganden ausbilden ($h_{VorlageApolar,i}$). Diese ist das Verhältnis der Anzahl apolarer Reste zur Anzahl aller Reste (n) an der Position *i*.

$$h_{VorlageApolar,i} = \frac{\text{Anzahl beobachteter apolarer Reste an Position } i}{n}$$

Zusätzlich wird die erwartete Häufigkeit für Reste ohne HB-Gruppe ($h_{ErwartetApolar,i}$) abgeschätzt. Diese ist die Häufigkeit, mit der apolare Reste an der Position gefunden würden, wenn die Reste gemäß der mittleren Häufigkeiten aus der Rotamerbibliothek verteilt wären. Dazu werden an der betrachteten Position sämtliche Rotamere aller Aminosäuren eingesetzt und es wird bestimmt, welche Rotamere keine HB-Gruppe zum Liganden ausbilden können. Für diese Rotamere werden die relativen Häufigkeiten aus der Rotamerbibliothek übernommen. $h_{ErwartetApolar,i}$ ist die Summe dieser relativen Häufigkeiten.

Werden $h_{VorlageApolar,i}$ und $h_{ErwartetApolar,i}$ als Wahrscheinlichkeiten interpretiert, so lässt sich das apolare Potential der Position *i* berechnen:

$$E_{Funktionsdef\ apolar,i} = -\ln\left(\frac{h_{VorlageApolar,i}}{h_{ErwartetApolar,i}}\right)$$

Das Potential der HB-Gruppen der Position *i* wird ähnlich ermittelt. Bei dieser Berechnung wird aber zusätzlich die räumliche Verteilung der HB-Gruppen bewertet. Für

die räumliche Verteilung werden alle HB-Gruppen als Menge von Raumpunkten zusammengefasst. Aus diesen Raumpunkten wird eine Wahrscheinlichkeitsdichte $g_{Vorlage,i}$ mit einer multivariaten Gaussfunktion geschätzt. Diese approximiert die räumliche Verteilung aller HB-Gruppen einer Position. Der Raum, in dem fast alle Punkte liegen, lässt sich durch einen Ellipsoid $ellipso_i$ beschreiben, der durch die zweifache Standardabweichung von $g_{Vorlage,i}$ gegeben ist. Mit $ellipso_i$ lässt sich eine zweite, Wahrscheinlichkeitsdichte $g_{Erwartet,i}$ beschreiben, die im Raumvolumen des Ellipsoids gleichverteilt ist. Durch die beiden Wahrscheinlichkeitsdichten kann nun beurteilen werden, ob die HB-Gruppen Präferenzen hinsichtlich der räumlichen Verteilung haben. Diese drückt sich im Verhältnis der beiden Wahrscheinlichkeitsdichten innerhalb von $ellipso_i$ aus.

Neben der räumlichen Präferenz wird auch die Konserviertheit der HB-Gruppe bestimmt. Hierfür werden wiederum die relativen Häufigkeiten ermittelt. Da die HB-Gruppen einer Position in der Strukturbibliothek von unterschiedlichen Aminosäuretypen stammen können, wird die Menge der Reste nach Aminosäuren aufgeteilt. So lassen sich relative Häufigkeiten pro Aminosäuretyp bestimmen ($h_{VorlageHB,i}$).

Zusätzlich werden die erwarteten Häufigkeiten der HB-Gruppe dieser Aminosäuretypen bestimmt. Dazu werden alle Rotamere dieser Aminosäuren an der betrachteten Position eingesetzt. Es werden dabei nur Rotamere gezählt, deren HB-Gruppe innerhalb des Ellipsoids $ellipso_i$ liegt. Die relative Häufigkeit für diese Rotamere lässt sich wiederum aus der Rotamerbibliothek ermitteln ($h_{ErwartetHB,i}$).

Damit das Potential für HB-Gruppen an der Position i definiert werden kann, werden die ermittelten Verteilungen $h_{VorlageHB,i}$ und $h_{ErwartetHB,i}$ als Wahrscheinlichkeiten interpretiert. Da angenommen wird, dass die Wahrscheinlichkeit für HB-Gruppen unabhängig ist von deren räumlichen Verteilung, werden die Wahrscheinlichkeitsverteilungen $h_{VorlageHB,i}$ und $h_{ErwartetHB,i}$ mit den räumlichen Wahrscheinlichkeitsdichten $g_{Vorlage,i}$ und $g_{Erwartet,i}$ als Produktwahrscheinlichkeit kombiniert. Mit diesen Produktwahrscheinlichkeiten lässt sich dann ein Potential definieren:

$$E_{Funktionsdef\ HB,i}(as,x) = -\ln\left(\frac{h_{VorlageHB,i}(as)\,g_{VorlageHB,i}(x)}{h_{ErwartetHB,i}(as)\,g_{ErwartetHB,i}(x)}\right)$$

Mit $E_{Funktionsdef\ HB,i}$ lässt sich nun ermitteln, wie gut die HB-Gruppe einer Aminosäure *as,* die sich räumlich an der Position x befindet, zu den HB-Gruppen an Position i in der Strukturbibliothek passt.

Die beiden Potentiale $E_{Funktionsdef\ HB,i}$ und $E_{Funktionsdef\ apolar,i}$ lassen sich nun zu einem Potential zusammenfassen:

$$E_{Funktionsdef,i}(as,x) = \begin{cases} E_{Funktionsdef\ HB,i}(as,x) & \textit{falls die HB - Gruppe der Aminosäure as} \\ & \textit{mit Raumposition x in elipso}_i \textit{ liegt} \\ E_{Funktionsdef\ apolar,i} & \textit{sonst} \end{cases}$$

Dabei werden zwei Fälle unterschieden: 1) Die HB-Gruppe der Aminosäure *as* mit den Raumkoordinaten *x* liegt in $ellipso_i$. In diesem Fall wird $E_{Funktionsdef\ HB,i}$ berechnet. 2) Die HB-Gruppe liegt außerhalb von $ellipso_i$. In diesem Fall wird $E_{Funktionsdef\ apolar,i}$ berechnet.

Obwohl die Potentiale positionsspezifisch aus Rotamerverteilungen abgeleitet worden sind, ist die Verwendung der Potentiale weder an Rotamere noch an Positionen gebunden. Die Potentiale abstrahieren von beidem, da sie nur noch HB-Gruppenverteilungen im Raum beschreiben. Mit ihnen lässt sich beurteilen, wie ähnlich die HB-Gruppen-Verteilung in dem aktiven Zentrum eines Proteinmodells den HB-Gruppen-Verteilungen in den Vorlagen der Strukturbibliothek ist.

4.1.3.4 Energieberechnung mit der Funktionsdefinition

Die oben eingeführte Funktionsdefinition umfasst eine Menge von Potentialen. Jedes Potential repräsentiert eine HB-Gruppen-Verteilung und kann bewerten, wie gut eine HB-Gruppe aus einem Modell zu dieser Verteilung passt.

Der einfachste Weg, die Potentiale mit der Energiefunktion des Proteindesigns zu verrechnen, wäre, die Energien pro Rotamer mit HB-Gruppe eines Modells zu bestimmen. Diese Energien könnten dann als Selbstenergie der Energietabelle hinzugefügt werden. Dadurch wäre dieses Modul ähnlich einfach in TRANSCENT integriert wie das Modul für die Ligandenbindung (vgl. 4.1.2.2). Allerdings hätte dieser Ansatz einen entscheidenden Nachteil: Die Optimierung könnte nicht erfassen, wie oft eine HB-Gruppe aus der Funktionsdefinition im Modell realisiert wird. Das Problem liegt darin, dass die Ähnlichkeit als Energie kodiert wäre und dass der Energietabelle damit nicht mehr zu entnehmen wäre, ob und welche HB-Gruppe mit einem Rotamer modelliert wird.

Basierend auf diesen Überlegungen wird in TRANSCENT die Ähnlichkeit zur Funktionsdefinition wie folgt beurteilt: Nach jedem Optimierungsschritt wird die Menge der modellierten HB-Gruppen mit der Menge der geforderten HB-Gruppen abgeglichen. Dazu werden sich entsprechende Gruppen einander zugeordnet. So kann bestimmt werden, welche Gruppen mehrfach bzw. gar nicht realisiert worden sind und es können zur Korrektur der Gesamtenergie entsprechende Energiebeiträge aufaddiert werden.

4.1.3.5 Das Zuordnungssystem

Für die Berechnung der Energie aus den Potentialen der Funktionsdefinition wird eine Zuordnung von HB-Gruppen zu Potentialen benötigt. Daher werden zuerst die Rotamere des Modells bestimmt, die eine HB-Gruppe in der Nähe des Liganden haben. Für jede dieser HB-Gruppen werden dann die Energien berechnet, die der HB-Gruppe in den einzelnen Potentialen der Funktionsdefinition zugewiesen werden. Jeder dieser Energiewerte beschreibt, wie gut die HB-Gruppe zu den HB-Gruppen-Verteilungen passt, die durch die dazugehörigen Potentiale repräsentiert werden. Die berechneten Werte werden dann als Potentialvektor gespeichert. Die Menge der Potentialvektoren für alle gefundenen Rotamere des Modells beschreiben, wie gut jede HB-Gruppe zu jedem Potential der Funktionsdefinition passt. Auf diese Weise sind die Energien für alle Kombinationen von Zuordnungen vorberechnet.

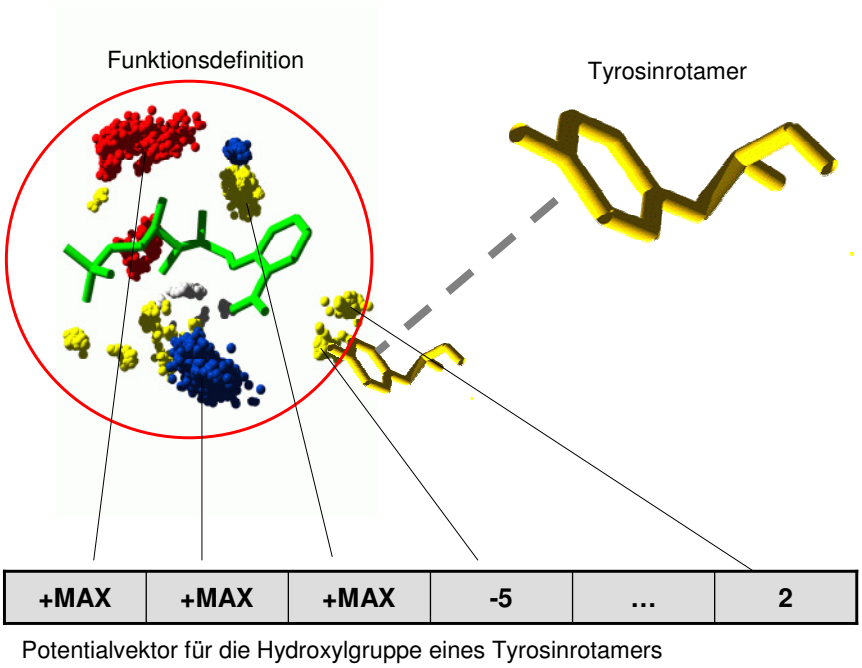


Abbildung 15: Berechnung eines Potentialvektors
Um die Ähnlichkeit zwischen Modell und einer Funktionsdefinition zu bestimmen, wird für jedes Rotamer, das eine HB-Gruppe in der Nähe des Liganden positioniert, ein Potentialvektor berechnet. Darin wird pro Potential der Funktionsdefinition ein Energiewert gespeichert, der beschreibt, wie gut die betrachtete HB-Gruppe in das jeweilige Potential passt. In der Abbildung wird dieser Ablauf durch ein Beispiel veranschaulicht: Die Potentiale der Funktionsdefinition (roter Kreis) werden durch Punktwolken repräsentiert. Die Hydroxylgruppe eines Tyrosinrotamers liegt nahe am Liganden. Darum wird für alle Potentiale die Energie des Rotamers berechnet. Die Übereinstimmung des Rotamers mit zwei räumlich nahen Potentialen drückt sich in niedriger Energie aus (-5, 2). Alle anderen Potentiale sind nicht mit dem Rotamer vereinbar, was sich in einer hohen Energie ausdrückt (+MAX).

Die Menge der gefundenen Vektoren beschreibt ein Zuordnungsproblem *Zuord(HB, POT)*. Gesucht ist eine paarweise Zuordnung von HB-Gruppen (*HB*) zu Potentialen (*POT*), so dass die Summe der damit verbundenen Energiewerte minimal wird. *Zuord* wird hier mit der so genannten *ungarischen Methode* (Kuhn, 1955) gelöst. Das Verfahren liefert eine optimale Zuordnung bei einer Laufzeit von $O(n^3)$, wobei *n* hier für die Anzahl der Potentiale steht.

Mit der Lösung für *Zuord* lässt sich die Gesamtenergie für die Funktionsdefiniton berechnen. Dazu müssen drei Fälle unterschieden werden: 1) Für die einander zugeordneten Paare lässt sich der Energiebeitrag über den Vektor der HB-Gruppe bestimmen. 2) Wird eine HB-Gruppe nicht zugeordnet, so wird ein Strafenergiebeitrag addiert, da eine nicht zugeordnete HB-Gruppe bedeutet, dass sie der Vorlage überhaupt nicht entspricht. 3) Wird dagegen ein Potential nicht zugeordnet, dann wird der Energiewert für die hydrophobe Ausprägung des Potentials addiert, da ein nicht zugeordnetes Potential bedeutet, dass das Proteinmodell in dem Bereich des Potentials hydrophob ist.

Dieses Modul hat zwei Effekte auf die Optimierung: Zum einen werden HB-Gruppen an die Stellen positioniert, die möglichst gut der Funktionsdefinition entsprechen. Zum anderen werden HB-Gruppen an solchen Stellen unterdrückt, an denen in den Vorlagen keine HB-Gruppen beobachtet wurden. Während sich der erste Effekt aus dem Konzept der Potentiale ergibt, ist der zweite Effekt indirekt. Wird eine HB-Gruppe an einer Stelle im Raum positioniert, an der sie keinem der HB-Gruppe Merkmale entspricht, bleibt die Zuordnung des dazugehörigen Rotamers erfolglos und die Strafenergie wird aufgeschlagen.

In TRANSCENT wird dieses Modul mit den beiden ersten kombiniert und kann so neben Stabilität und Ligandenbindung auch die Ähnlichkeit zu aktiven Zentren der Vorlagestrukturen optimieren.

4.1.4 Modul 4: pKa-Wert-Optimierung

Für katalytisch essentielle Reste ist es nicht nur wichtig, dass sie die richtige Orientierung zum Liganden haben, sondern auch, dass ihre pKa-Werte im richtigen Bereich liegen. Diese pKa-Werte lassen sich durch benachbarte Reste beeinflussen und optimieren. In TRANSCENT werden daher beim Modellieren eines aktiven Zentrums auch die pKa-Werte berücksichtigt. Dazu werden Referenz-pKa-Werte aus den Beispielstrukturen berechnet und als weitere Bedingung für die Optimierung verwendet. Diese pKa-Optimierung bildet das vierte Modul.

4.1.4.1 Geschwindigkeitsoptimierung der PROPKA-Methode

Für die Berechnung der pKa-Werte wird die PROPKA-Methode verwendet. Trotz der relativ guten Performanz ist die Originalversion von PROPKA für die Aufgabe im Proteindesign nicht nutzbar. Die Auswertung eines Proteinmodells dauert mit PROPKA etwa eine Sekunde. Für das verwendete SA-Protokoll müssen oft mehrere Millionen von Optimierungsschritten durchgeführt werden. Für die pKa-Wert-Optimierung würden entsprechend viele Aufrufe des PROPKA-Programms benötigt. Damit läge der Rechenaufwand für diesen Optimierungsprozess in der Größenordnung von Wochen. Es war daher nötig, die Laufzeit der PROPKA-Methode zu optimiert.

Gleichzeitig wurde die Methode an das Proteindesignkonzept angepasst. Das zu Grunde liegende Konzept ist eine rotamerorientierte Dekomposition der pKa-Wert-Verschiebungen, analog zum Konzept der rotamerorientieren Dekomposition der Energie im Proteindesign. Hierfür werden alle pKa-Wert-Verschiebungen, die sich zwischen zwei Rotameren ergeben können, vorberechnet und in einer Tabelle abgespeichert (vgl. Kapitel 2.3.1.4). Dabei werden die statischen Effekte separat von den dynamischen Effekten gespeichert, da diese skaliert werden müssen und zwar abhängig vom Protonierungszustand des gesamten Proteins (vgl. Kapitel 2.3.4.4).

Auf diese Weise vorberechnet, lassen sich für die Rotamere eines Proteinmodells die pKa-Wert-Verschiebungen während der Optimierung nachschlagen. Dadurch entfallen aufwändige Mehrfachberechnungen. Die statischen Effekte lassen sich direkt zusammenfassen. Für die dynamischen Effekte wird weiterhin die rundenbasierte Optimierung des Protonierungszustands durchgeführt (vgl. Kapitel 2.3.4.4).

Beim SA ändert sich das Proteinmodell nur um ein Rotamer pro Optimierungsschritt, so dass große Teile des Modells unverändert bleiben. Daher werden mit Hilfe eine Updatefunktion nur die differenziellen Änderungen angepasst und weite Teile des vorherigen Berechnungsergebnisses können beibehalten werden (Leaver-Fay et al., 2007).

Mit der Updatefunktion wird dabei zunächst der Einfluss des zu verändernden Rotamers auf die pKa-Werte bestimmt und abgezogen. Anschließend wird analog der Effekt des neuen Rotamers addiert. Der Protonierungszustand wird nur reoptimiert, wenn der Tausch des Rotamers zu einer Änderung der dynamischen Effekte geführt hat. Auf diese Weise wird der Rechenaufwand für die pKa-Wert-Optimierung nochmals verringert.

Die Methode kann von dieser Optimierung des Rechenaufwands nur im Rahmen des Proteindesignprozesses profitieren. Werden nur die pKa-Werte für ein einzelnes Proteinmodell berechnet, so ist keine der beschriebenen Rechenzeitoptimierungen anwendbar. Für eine pKa-Wert-Optimierung ist das Programm bei identischer Präzision etwa 5000-mal schneller als die Originalversion. Insgesamt liegt der Rechenaufwand für

eine komplette Optimierung nun in der Größenordnung von Minuten. Daher kann die PROPKA-Methode für die Umwandlungsmodellierung verwendet werden.

4.1.4.2 Kopplung der pKa-Wert-Optimierung mit der Funktionsdefinition

Die Menge der katalytisch essentiellen Reste wird durch die Funktionsdefinition beschrieben. Erst im Verlauf der Optimierung wird diese an einer geeigneten Stelle im neuen aktiven Zentrum positioniert. Damit besteht für die pKa-Wert-Optimierung ein Zuordnungsproblem: Für die Entscheidung, ob eine pKa-Optimierung erforderlich ist, muss bestimmt werden, ob die Reste bereits positioniert wurden und gegebenenfalls wo sie positioniert wurden (Abbildung 16).

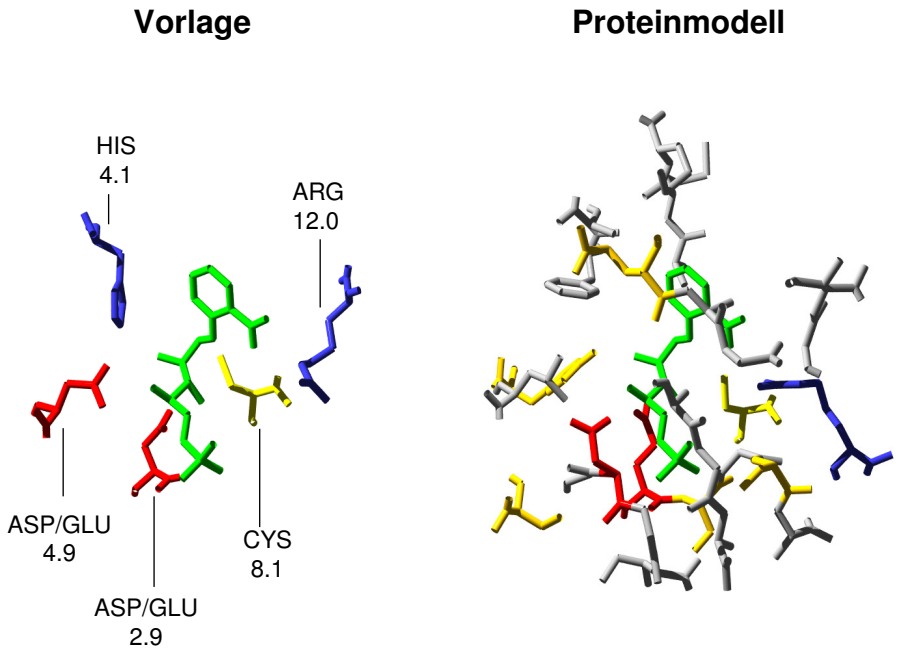


Abbildung 16: Zuordnungsproblem bei der pKa-Wert Optimierung
Links sind fünf Reste des aktiven Zentrums von TrpF zu sehen, für die Referenz-pKa-Werte vorgegeben sind. Rechts ist ein Proteinmodell zu sehen, auf welches das aktive Zentrum übertragen werden soll. Die pKa-Werte können in diesem Modell nicht direkt angepasst werden, da ein Zuordnungsproblem besteht: Zunächst müssen die korrespondierenden Reste zwischen Vorlage und Modell bestimmt werden. Anschließend kann entschieden werden ob eine pKa-Optimierung im Modell erforderlich ist.
Die Färbung zeigt Arg, His sowie Lys in blau. Asp und Glu sind rot eingefärbt. Asn, Cys, Gln, Ser, Thr sowie Tyr sind gelb und der Ligand ist grün eingefärbt. Die hydrophoben Reste sind grau dargestellt.

Das Zuordnungssystem des Moduls für die Ähnlichkeitsoptimierung (vgl. 4.1.3.5) löst dieses Problem bereits für die Beurteilung der HB-Gruppen. Da sich erst mit der

Zuordnung entscheiden lässt, welcher der pKa-Werte eines aktiven Zentrums optimiert werden muss, liegt es nahe, das pKa-Modul an das Zuordnungssystem zu koppeln (Abbildung 17).

Für die Optimierung wird die Kopplung wie folgt realisiert: Alle Reste der Vorlage, deren pKa-Wert als Referenz-pKa-Wert für die Optimierung definiert wurde, sind in der Funktionsdefinition mit einem Potential vertreten. Somit werden die Referenz-pKa-Werte mit den Potentialen assoziiert.

Um die pKa-Werte eines Modells während der Optimierung zu beurteilen, wird das pKa-Modul stets nach dem Funktionsdefinitionsmodul aufgerufen, um so die Zuordnung des Funktionsdefinitionsmoduls für die pKa-Optimierung verwenden zu können. Das Funktionsdefinitionsmodul identifiziert alle Reste, die mit Potentialen aus der Funktionsdefinition korrespondieren und ordnet diese einander zu. Nachdem es mit der Zuordnung die Ähnlichkeit berechnet hat, übergibt es die Zuordnungsinformation an das pKa-Modul. Dieses berechnet zunächst alle pKa-Werte des Modells. Anschließend ermittelt es mit Hilfe der Zuordnungsinformation, welche Reste im Modell einem Rest mit Referenz-pKa-Wert in der Vorlage zugeordnet worden sind. Für diese *n* Paare berechnet es den Unterschied der pKa-Werte mit folgendem Strafterm (vgl. Kapitel 2.3.1.2):

$$E_{PROPKA} = wF \sum_{i=1}^{n} \left(\left| pKa_{Referenz,i} - pKa_{Modell,i} \right| \right)^{wP}$$

Hierbei stellen *wF* und *wP* Gewichte für den Strafterm dar. Für die verbleibenden Referenz-pKa-Werte, für die in der Zuordnung kein korrespondierender Rest im Modell aufgeführt ist, wird ein maximaler Unterschied von 7 pH-Einheiten angenommen. Auf diese Weise ist die Abweichung des Modells von allen Referenz-pKa-Werten bestimmt.

Aus den Unterschieden wird schließlich unter Verwendung eines harmonischen Strafterms eine Abweichungsenergie berechnet. Je größer die Unterschiede zwischen Modell und Referenz, desto höher wird das Modell energetisch bewertet. Auf diese Weise wird während der Optimierung versucht, solche Reste zu etablieren, die den Resten mit den Referenz-pKa-Werten zugeordnet werden können. Zudem wird versucht, die pKa-Werte dieser korrespondierenden Reste an die Referenzwerte anzugleichen.

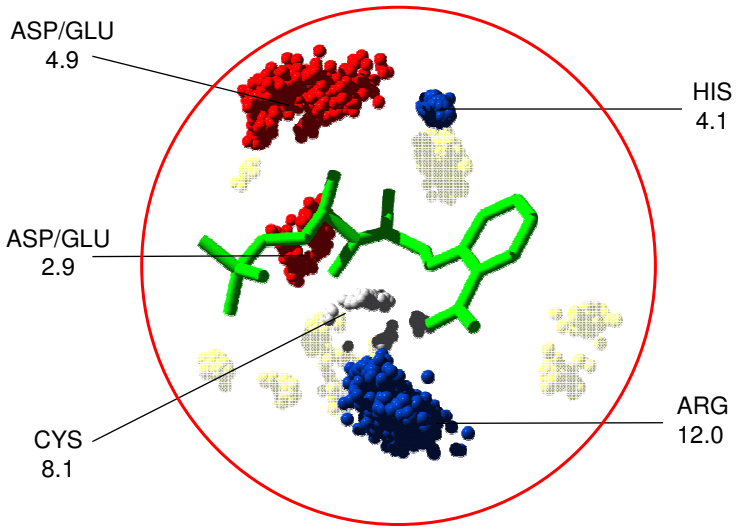


Abbildung 17: Die Kopplung von Funktionsdefinition und pKa-Werten
Aus den Vorlagestrukturen werden nicht nur Potentiale, sondern auch Referenz-pKa-Werte abgeleitet. Diese geben den pKa-Wert für ein Rotamer vor, das einem Potential zugeordnet worden ist. In der Abbildung sind die Potentiale der TrpF Funktionsdefinition durch Punktwolken angedeutet. Fünf dieser Potentiale fordern titrierbare Reste. Daher sind sie jeweils mit einem Referenz-pKa-Wert assoziiert.

Das pKa-Berechnungsmodul wird der Energiefunktion von TRANSCENT als Strafterm hinzugefügt. Auf diese Weise wird das Modul mit den drei anderen Modulen kombiniert. So kann TRANSCENT neben Stabilität, Ligandenbindung und Ähnlichkeit zum aktiven Zentrum der Vorlage zusätzlich die pKa-Werte von ligandennahen Resten optimieren.

4.1.5 Zusammenführung der Module

Wie oben dargestellt, enthält TRANSCENT vier Module. Jedes der Module stellt dabei die Wahrung einer anderen Rahmenbedingung sicher. Jedes Modul kann für ein Proteinmodell berechnen, zu welchem Grad es die jeweilige Rahmenbedingung erfüllt ist. Hierbei wird der Grad der Anpassung in Form eines Energieterms ausgedrückt. Damit die Module den Optimierungsprozess beeinflussen können, werden deren Energiebeiträge in einer Energiefunktion zusammengeführt:

$$E_{TRANSCENT} = w_1 E_{ROSETTA} + w_2 E_{DRUGSCORE} + w_3 E_{Funktionsdef} + w_4 E_{PROPKA}$$

Die dafür notwendige Gewichtung w_1 … w_4 wird im folgenden Kapitel 4.3 im Detail beschrieben. Da die Gesamtenergie für jeden Optimierungsschritt neu berechnet wird, nehmen alle Module gleichzeitig Einfluss auf den Verlauf der Optimierung.

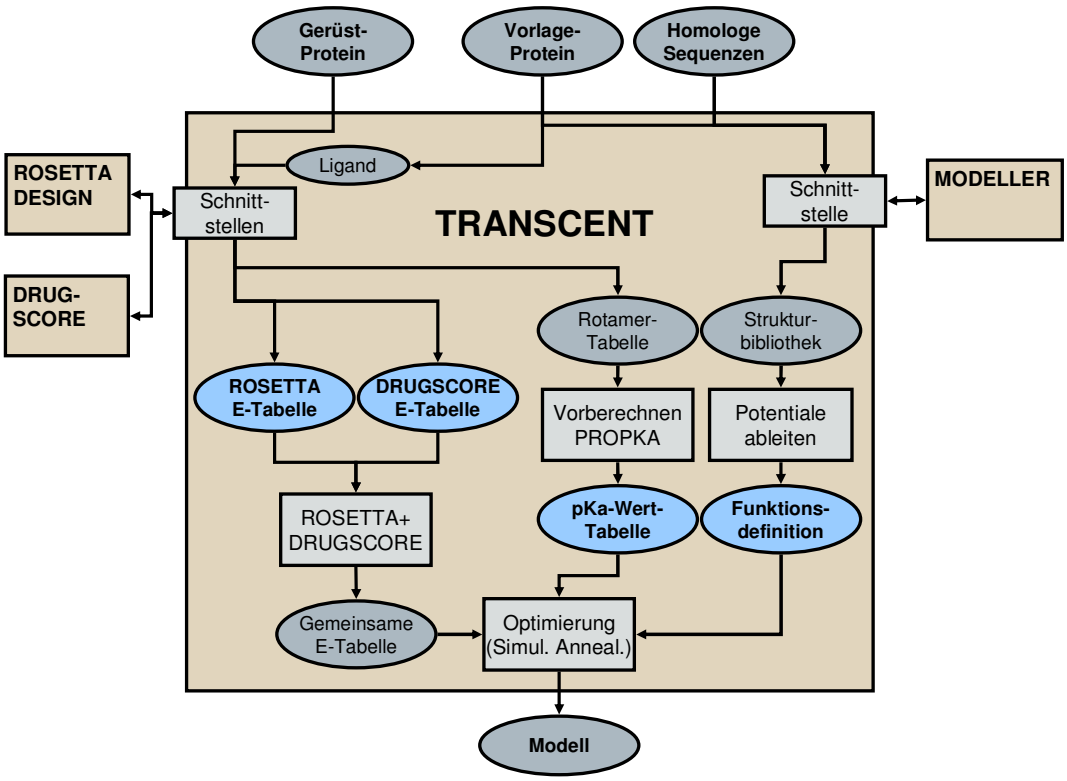


Abbildung 18: Schematischer Programmablauf in TRANSCENT

TRANSCENT erwartet für eine automatische Übertragung eines aktiven Zentrums folgende Eingaben: Die Struktur des Proteins, auf dem das neue aktive Zentrum etabliert werden soll, die Struktur des zu übertragenden aktiven Zentrums mit gebundenem Liganden und eine Menge von Sequenzen, die Homolog zur Vorlage sein müssen.

Mit diesen Eingaben werden vier Datenstrukturen (blau hinterlegte Ellipsen) erzeugt. Diese bilden die Grundlage der vier Module von TRANSCENT: Mit ROSETTA DESIGN und DRUGSCORE als externe Programme werden zwei Energietabellen berechnet. Mit der von ROSETTA DESIGN berechneten Rotamertabelle wird über die PROPKA-Methode eine Tabelle vorausberechnet, die Werte für pKa-Perturbationen enthält. Mit Hilfe der Vorlagestruktur und der homologen Sequenzen berechnet MODELLER die Strukturbibliothek, aus der die Funktionsdefinition abgeleitet wird. Auf Grundlage dieser Datenstrukturen erstellt das Optimierungsverfahren ein Modell mit neuem aktiven Zentrum erstellt. Dieses wird schließlich ausgegeben.

Mit der hier vorgestellten Kombination von vier Modulen kann TRANSCENT den Transfer von aktiven Zentren vollständig und automatisiert modellieren (Abbildung 18). Der modulare Aufbau von TRANSCENT stellt zum einen eine einfache Erweiterbarkeit sicher, zum anderen erlaubt er verschiedene Kombinationen der Module für die Modellierung zu verwenden

4.2 Untersuchung der Qualität von Homologiemodellen

Für die vorliegende Arbeit ist die Qualität der Homologiemodelle von großer Bedeutung, da aus ihnen die Funktionsdefinition abgeleitet wird (vgl. Kapitel 4.1.3). Beim Homologiemodellieren wird für eine Eingabesequenz ein Strukturmodell erzeugt, wobei homologe Strukturen als Vorlage dienen. In einfachen Fällen sind große Teile der Eingabe mit der Sequenz einer Vorlagestruktur identisch. Dann können entsprechende Teile der Vorlagestruktur übernommen werden. Das Modellieren beschränkt sich in diesen Fällen vor allem auf das Modellieren der Unterschiede. Aus diesem Grund ist die Homologiemodellierung zurzeit die erfolgreichste Methode zur Strukturvorhersage (Ginalski, 2006). Wie oben erläutert, wird für eine Umwandlungsmodellierung mit TRANSCENT eine Menge von Strukturbeispielen für das aktive Zentrum des Vorlageproteins benötigt. Um die Datenbasis zu vergrößern (vgl. Kapitel 4.1.3), werden Strukturmodelle von aktiven Zentren aus homologen Sequenzen, die aus anderen Organismen stammen, mittels Homologiemodellierung generiert. Homologiemodellierungen mit nahe verwandten Proteinen sind besonders einfach, da in der Regel eine hohe Sequenzidentität zwischen Eingabesequenz und Strukturvorlage besteht. Von den modellierten Strukturen sind anschließend nur die aktiven Zentren von Interesse. Aktive Zentren sind oft besonders konserviert, was die Modellierung zusätzlich vereinfacht.

Im betrachteten Fall wird also die erfolgreichste Methode (Homologiemodellierung) für eine einfache Problemstellung (gleiche Proteine aus verschieden Organismen) verwendet und nur ein gut konservierter Teil der Struktur (aktives Zentrum) ist relevant. Daher ist eine hohe Modellierungsqualität zu erwarten. Trotzdem ist es nötig, die Modellierungsqualität quantitativ zu erfassen, um die Befunde abzusichern, die aus den vorhergesagten Modellen abgeleitet werden.

In diesem Kapitel werden Fehlerquellen, die bei der Homologiemodellierung auftreten können, genauer untersucht. Dazu wird ein Testdatensatz vorgestellt, mit dem anschließend verschiedene Modellierungsstrategien beurteilt werden. Außerdem wird ein Kriterium beschrieben, mit dem die Modellierungsqualität beurteilt werden kann.

4.2.1 Ein strukturunabhängiges Maß für die Modellierungsqualität

Um die Qualität eines Strukturmodells zu erfassen, muss die Abweichung zwischen Modell und einer Referenzstruktur berechnet werden. Dazu wird üblicherweise die Wurzel der mittleren quadratischen Abweichung (RMSD) zwischen äquivalenten Atompositionen beider Strukturen berechnet (Carugo & Pongor, 2001). Im hier betrachteten Anwendungsfall lässt sich die Modellierungsqualität auf diese Weise jedoch nicht erfassen, da keine Referenzstruktur zur Verfügung steht. Ansonsten wäre keine

Modellierung notwendig. Die Erfassung der Modellierungsqualität durch Berechnung des RMSD-Werts ist also nur bei bekannter Referenzstruktur in Form von Testmodellierungen möglich.

Um die Modellierungsqualität auch im hier betrachteten Anwendungsfall abschätzen zu können, wird ein strukturunabhängiges Kriterium notwendig, das es erlaubt die Modellierungsqualität auch ohne Referenz abzuschätzen. Ein derartiges Kriterium ist die Beurteilung der Sequenzidentität, wie es zum Beispiel das Programm MODELLER verwendet (Eswar et al., 2007; Sali & Blundell, 1993). Hierfür wird der Grad der Sequenzidentität zwischen Templat und Modellsequenz bestimmt. Liegt die Sequenzidentität über 30%, so wird für MODELLER eine durchschnittliche Modellierungsqualität mit einem RMSD von unter 3.5Å angegeben.

Diese Abschätzung ist jedoch für diese Arbeit aus mehreren Gründen nicht ausreichend. Erstens ist der Wert nur für die gesamte Proteinstruktur bestimmt worden. Da in der Strukturbibliothek aber vor allem die aktiven Zentren von Bedeutung sind, ist ein gesondert berechneter RMSD-Wert für aktive Zentren wichtig. Außerdem wäre ein mittlerer RMSD von 3.5Å für aktive Zentren zu hoch. Da für die Katalyse die richtige geometrische Anordnung der aktiven Seitenketten entscheidend ist, sollte der durchschnittliche RMSD-Wert kleiner als 2Å sein. Ein weiteres Problem ist, dass der Wert nur für Proteinrückgratatome bestimmt wurde. Für TRANSCENT ist aber eine hohe Modellierungsqualität auch für die Atome der Seitenketten wichtig. Daher sollte der RMSD über alle Atome berechnet werden, um die Modellierungsqualität zu beurteilen.

Es bleibt zu klären, ob bei Modellen mit einer Sequenzidentität von mehr als 30% zum Templat der mittlere RMSD-Wert für alle Atome des aktiven Zentrums unter 2Å liegt.

4.2.2 Der Homologiemodellierungs-Testdatensatz

Soll ein Strukturvorhersageprogramm getestet werden, müssen die zu modellierenden Strukturen bekannt sein. In diesem Fall lassen sich Vorhersagen und reale Strukturen miteinander vergleichen, wodurch eine Beurteilung der Vorhersagequalität ermöglicht wird. Auf die beschriebene Weise werden beim CASP-Wettbewerb Strukturvorhersage-Programme beurteilt (Moult, 2005). Ein Testdatensatz für die Strukturvorhersage besteht also aus Proteinsequenzen und den dazu gehörigen Strukturen. Die Strukturen werden nur zum anschließenden Beurteilen verwendet.

Für die Test-Homologiemodellierungen müssen im Datensatz zu jedem Protein mindestens zwei Strukturen vorhanden sein. Zum einen muss eine Referenzstruktur für die Sequenz, die modelliert werden soll, verfügbar sein. Nur so lässt sich die Modellierungsqualität beurteilen. Zum anderen muss aber zusätzlich eine Struktur vorhanden sein, die als Templat für die Homologiemodellierung dient. Die Sequenz des Templats muss

sich von der zu modellierenden Sequenz unterscheiden, da ansonsten schon das Templat ein perfektes Strukturmodell ist. Als Templat kann jedoch kein anderes Protein mit gleicher Faltung verwendet werden, weil sich andernfalls der Testfall von der zu untersuchenden Fragestellung zu sehr unterscheidet. Hier wird ja vorausgesetzt, dass die Templatstruktur auch eine Vorlage für das aktive Zentrum liefert. Folglich werden mindestens zwei Strukturen des gleichen Proteins aus verschiedenen Organismen gebraucht (Abbildung 19).

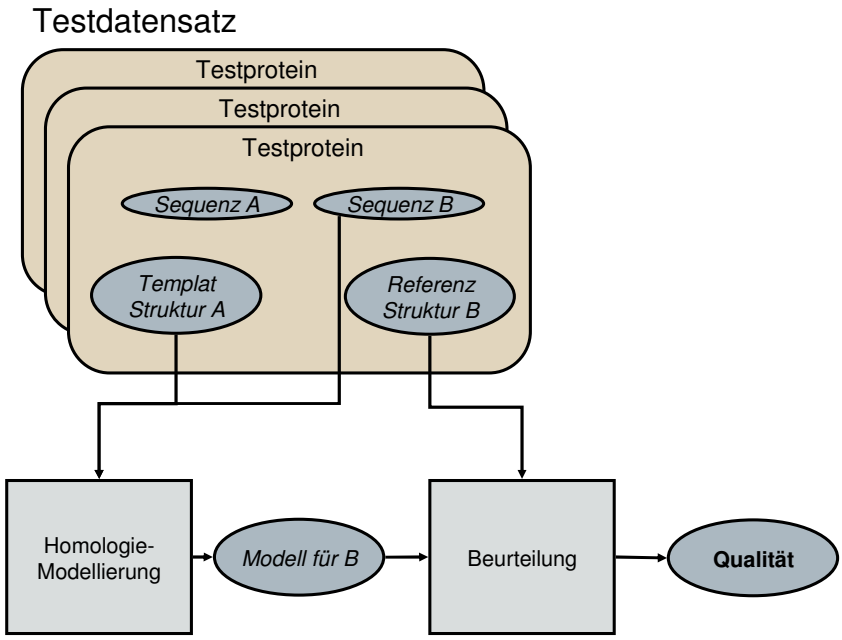


Abbildung 19: Beurteilung der Modellierungsqualität mit dem Testdatensatz
Jedes Element des Testdatensatzes besteht aus einem Paar von Strukturen (Struktur A und B) und deren Sequenzen (Sequenz A und B). Sequenz B wird mit Struktur A als Templat modelliert. Das entstandene Modell wird anschließend mit der Referenzstruktur B für Sequenz B verglichen.

Der Testdatensatz muss zudem die Information enthalten, wo sich das aktive Zentrum befindet. Nur so lässt sich die Modellierungsqualität für diese Region separat bewertet. Zur Bestimmung der aktiven Zentren kann wie folgt vorgegangen werden:

Ist in einer Struktur im aktiven Zentrum ein Ligand eingebettet, dann lassen sich die Reste des aktiven Zentrums mit einem Abstandskriterium (7Å um den Liganden) bestimmen. Es genügt, dass in jeweils einer Struktur des Proteins ein Ligand vorhanden ist, da aktive Zentren weitgehend konserviert sind. Die Positionen lassen sich mit einem Alignment auf andere Strukturen ohne Ligand übertragen.

4.2.3 Der Ribulosephosphat- $(\beta\alpha)_8$-Barrel Datensatz

Die Vertreter der Ribulosephosphat-$(\beta\alpha)_8$-Barrel Superfamilie eignen sich in besonderer Weise als Testdatensatz zur Beurteilung der Modellierungsqualität, da TRANSCENT im Folgenden mit Enzymen der Ribulosephosphat-$(\beta\alpha)_8$-Barrel Superfamilie evaluiert wird (vgl. Kapitel 2.2.2). Somit ist eine Beurteilung der Vorhersagequalität für diese Enzymklasse von besonderem Interesse.

Ganz generell stellen diese Enzyme einen nahezu idealen Testdatensatz dar, da sie umfassend charakterisiert sind. Zu jedem Protein sind mindestes zwei Strukturen aus verschiedenen Organismen bekannt. Aufgrund der Existenz eines Katalysepols (vgl. Abbildung 4) sind die aktiven Zentren leicht zu bestimmen. Weiterhin ist für nahezu jedes Protein eine Struktur mit Ligand bekannt, somit ist auch die Ermittlung der einzelnen Reste des jeweiligen aktiven Zentrums einfach. Zusätzlich sind für alle Barrelproteine viele homologe Sequenzen vorhanden, die in Form eines MSAs ebenfalls wichtig für den Testdatensatz sind (siehe Kapitel 4.2.4).

Aus diesen Gründen wurde ein Testdatensatz, bestehend aus Vertretern der Ribulosephosphat-$(\beta\alpha)_8$-Barrel Superfamilie zusammengestellt. Dieser erlaubt es, die Modellierungsqualität von Strukturmodellen zu beurteilen. Für den Testdatensatz wurden die folgenden Barrel-Proteine ausgewählt: HisA, HisF, TrpA, TrpC, TrpF, RP Epimerase (RPE) und OMP Decarboxylase (OMPD). Falls für ein Protein desselben Organismus mehrere Strukturen zur Verfügung standen, wurde die Wildtypstruktur einer Mutante vorgezogen. Die genaue Zusammensetzung des Testdatensatzes ist in Tabelle 3 gezeigt.

Zu HisA gibt es nur zwei Strukturen, zu TrpF und RPE gibt es drei Strukturen und zu den übrigen Proteinen jeweils vier Strukturen. Dieser Datensatz dient als Grundlage für die Definition von Testmodellierungen. Testmodellierungen sind Homologiemodellierungen bei denen die Modellstruktur mit einer Referenzstruktur verglichen werden kann. Die Anzahl der Testmodellierungen ergibt sich, indem für jedes Protein jede Struktur einmal als Templat für alle anderen dient. Auch triviale Modellierungen, bei denen die Referenzstrukturstruktur als Templat verwendet wird, werden betrachtet. Denn diese Modellierungen erlauben es, den Fehler abzuschätzen, den ein Modellierungsprogramm macht, wenn ideale Templatstrukturen zur Verfügung stehen. Berücksichtigt man diese Unterscheidung, so ergeben sich für HisA 4 Fälle (2 trivial), für TrpF und RPE 9 Fälle (3 trivial) und für die übrigen Proteine jeweils 16 Fälle (4 trivial). Insgesamt ermöglicht der Testdatensatz 86 Modellierungen, von denen 24 Fälle nicht trivial sind.

Tabelle 3: PDB-Codes der Strukturen im Ribulosephosphat- $(\beta\alpha)_8$Barrel-Testdatensatz
In der Tabelle sind die Proteine aus dem Testdatensatz aufgelistet, mit dem ein strukturunabhängiges Gütekriterium zur Beurteilung von Homologiemodellen ermittelt werden soll. Zu jeder Struktur sind Organismus und PDB-Code aufgelistet.

Protein	Organismus	PDB-Code
HisA	*Thermotoga maritima*	1qo2
	Mycobacterium tuberculosis	1vzw
HisF	*Thermotoga maritima*	1thf
	Thermus thermophilus	1ka9
	Saccharomyces cerevisiae (Hefe)	1ox5
	Pyrobaculum aerophilum	1h5y
TrpA	*Salmonella typhimurium*	1ttq
	Thermus thermophilus	1ujp
	Pyrococcus furiosus	1geq
	Escherichia coli	1x7y
TrpC	*Escherichia coli*	1pii
	Sulfolobus solfataricus	1igs
	Thermotoga maritima	1j5t
	Thermus thermophilus	1vc4
TrpF	*Escherichia coli*	1pii
	Thermotoga maritima	1lbm
	Thermus thermophilus	1v5x
RPE	*Solanum tuberosum* (Kartoffel)	1rpx
	Oryza sativa (Reis)	1h1z
	Synechocystis sp. strain PCC 6803	1tqj
OMPD	*Bacillus subtilis*	1dbt
	Escherichia coli	1eix
	Methanobacterium thermoautotrophicum	1lor
	Saccharomyces cerevisiae (Hefe)	1dqw

Die Sequenzidentität dieser 62 Fälle deckt ein Spektrum von 19% bis 61% ab. In Tabelle 4 sind die einzelnen Ähnlichkeitswerte aufgetragen. Die wechselseitigen Modellierungen zweier Strukturen stellen dabei nicht äquivalente Fälle dar, obwohl sie wechselseitig die gleiche Sequenzidentität aufweisen. Einleuchtend wird die Asymmetrie wenn z.B. die Sequenz zur Struktur *A* gegenüber der Sequenz zur Struktur *B* eine Insertion hat. Dann modelliert man im Fall „Sequenz zur Struktur *A* auf Struktur *B*“ eine Insertion, im Fall „Sequenz zur Struktur *B* auf Struktur *A*“ aber eine Deletion. Da Insertionen und

Deletionen Probleme sind, die mit völlig verschiedenen Ansätzen modelliert werden müssen, sind die Fälle nicht äquivalent.

Tabelle 4: Sequenzidentitäten der Proteine im Testdatensatz

In den Tabellen sind die Strukturen der Proteine des Testdatensatzes einander gegenübergestellt (PDB-Code). Die Zahlen der Tabelle geben die Sequenzidentität zwischen einzelnen Proteinpaaren in Prozent an. Jedes dieser Paare entspricht einer Testmodellierung des Testdatensatzes.

OMPD	1dbt	1dqw	1eix	1lor
1dbt	100	20	40	23
1dqw	20	100	19	20
1eix	40	19	100	23
1lor	23	20	23	100

TrpC	1igs	1j5t	1pii	1vc4
1igs	100	32	32	36
1j5t	32	100	31	33
1pii	32	31	100	35
1vc4	36	33	35	100

HisF	1h5y	1ka9	1ox5	1thf
1h5y	100	61	44	53
1ka9	61	100	44	58
1ox5	44	44	100	45
1thf	53	58	45	100

TrpA	1geq	1ttq	1ujp	1v7y
1geq	100	32	38	30
1ttq	32	100	26	84
1ujp	38	26	100	24
1v7y	30	84	24	100

TrpF	1lbm	1pii	1v5x
1lbm	100	30	39
1pii	30	100	27
1v5x	39	27	100

RPE	1h1z	1rpx	1tqj
1h1z	100	41	42
1rpx	41	100	67
1tqj	42	67	100

HisA	1qo2	1vzw
1qo2	100	23
1vzw	23	100

4.2.4 Der Vergleich von paarweisem und multiplem Sequenzalignment

Der erste Schritt bei der Homologiemodellierung ist das Alignment der Aminosäuren aus der zu modellierenden Sequenz und der Sequenz aus der Templatstruktur. Im Programm MODELLER wird dieses Alignment durch das Unterprogramm SALIGN automatisch durchgeführt. Das Programm erzeugt dazu ein paarweises Sequenzalignment (PSA). Eine alternative Vorgehensweise ist das Ableiten des benötigten PSAs aus einem MSA. Dazu werden zu den beiden Sequenzen weitere homologe Sequenzen in einem MSA aligniert. Anschließend werden die beiden Sequenzen als PSA aus dem MSA unter Beibehaltung der Zuordnung extrahiert. Im Kontext des MSA lassen sich die Sequenzen dann genauer alignieren (Marti-Renom et al., 2004).

Um den spezifischen Einfluss der Zuordnungsverfahren PSA und MSA auf die Modellierungsqualität zu beurteilen, ist ein ideales Alignment als Referenz notwendig. Da im Testdatensatz die Vorlagestruktur und die Referenzstruktur vorhanden sind, kann das

ideale Alignment mit Hilfe eines strukturbasierten, paarweisen Sequenzalignment (SPSA) generiert werden. In Abbildung 20 ist der Vergleich als Flussdiagramm veranschaulicht.

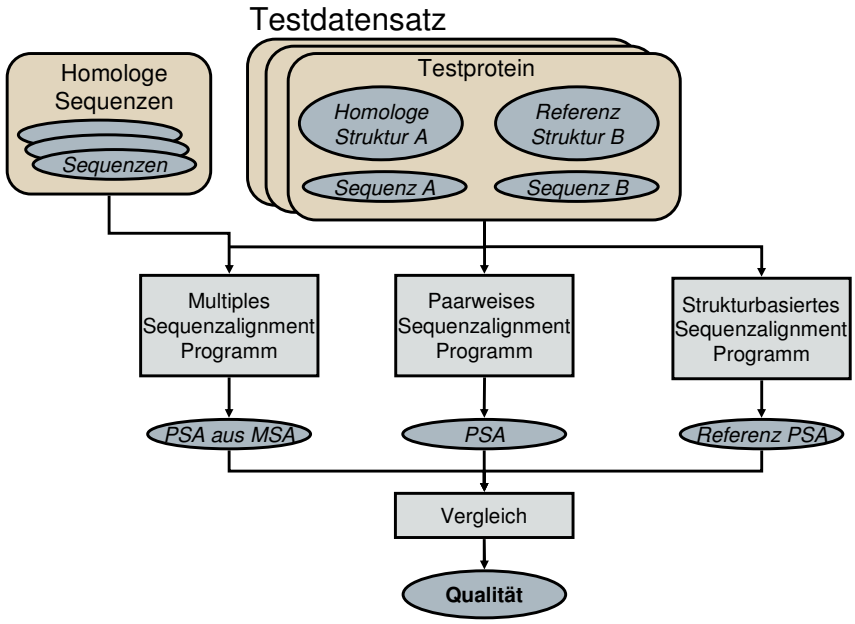


Abbildung 20: Beurteilung des Unterschieds von Zuordnungen mit PSAs und mit MSAs
Das Diagramm zeigt, wie die Sequenzen einzelner Testproteine als PSA und als PSA abgeleitet aus einem MSA aligniert werden. Um die Unterschiede beurteilen zu können, wird als Referenz ein strukturbasiertes PSA auf Basis der Strukturen des Testproteins erstellt.

Auch wenn allgemein bekannt ist, dass PSAs aus MSAs von besserer Qualität sind (Gribskov et al., 1987), ist nicht unmittelbar klar, welche Vorteile sich für die Regionen der aktiven Zentren ergeben. Hier ist die Konserviertheit höher als über das ganze Protein betrachtet. Generell verbessert eine höhere Konserviertheit die Qualität von paarweisen Sequenzalignments, da die Zuordnung dann einfacher wird.

4.2.5 Beurteilung von Modellierungen, die auf PSAs oder MSAs basierten

Bei der Zuordnung von Sequenzpositionen zu Strukturpositionen stehen generell die zwei Vorgehensweisen PSA-basiert und MSA-basiert zur Auswahl. Um die Zuordnungsqualität der Verfahren gegeneinander abzuwägen, wurde diese jeweils mit dem Barrel-Testdatensatz bewertet. Dazu wurden die Zuordnungsfehler bei den 62 nichttrivialen Testproteinen ermittelt (Abbildung 21).

Durch das PSA werden etwa 75% der Referenzzuweisungen gefunden, durch das MSA 84%. Dabei schwankt die Rate für die einzelnen Fälle erheblich. Die Zuordnungen des PSAs sind zwischen 16% und 100% identisch mit den Zuordnungen des

Referenzalignments, die des MSAs zwischen 49% und 100%. In fast allen Fällen ist die Zuordnung des PSAs schlechter als die des MSAs.

Beschränkt man sich bei der Auswertung auf die aktiven Zentren der Testproteine, ergibt sich folgendes Bild: Durch das PSA werden 78% der Referenzzuweisungen gefunden, durch das MSA 96%. Die Zuweisungsraten schwanken beim PSA zwischen 12% und 100% und beim MSA zwischen 78% und 100%. In allen Fällen ist die Zuordnung des PSAs schlechter als die des MSAs.

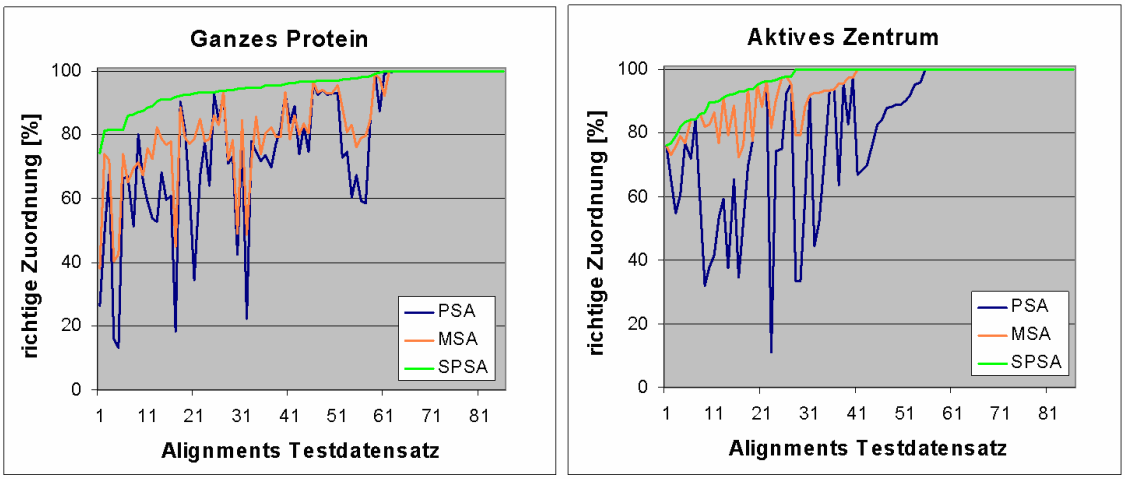


Abbildung 21: Vergleich der Zuordnungsfehler (Ganzes Protein / aktives Zentrum)
Die Diagramme (links: Ganze Proteine, rechts: Aktive Zentren) zeigen den Zuordnungsfehler für die Alignments der Proteinmodelle des Testdatensatzes. Die verglichenen Verfahren sind: Paarweises Sequenzalignment (PSA) in blau und das abgeleitete PSA aus einem MSA in orange. Sortiert sind die Alignments nach der maximal möglichen Zuordnung im strukturbasierten PSA (SPSA) in grün, das als Referenz dient.

Die obigen Tests belegen, dass ein MSA-basiertes Verfahren einen höheren Anteil korrekter Zuordnungen findet, als das PSA-basierte. Dieser Effekt ist im aktiven Zentrum besonders deutlich. Damit empfiehlt es sich, das MSA-basierte Verfahren zu verwenden. Allerdings kann das MSA-basierte Verfahren nur dann verwendet werden, wenn eine hinreichende Anzahl homologer Sequenzen verfügbar ist. Da die Homologiemodellierung in dieser Arbeit verwendet wird, um eine Strukturbibliothek zu erweitern, können die Sequenzen, auf denen die Strukturbibliothek basieren soll, auch für das MSA verwendet werden. Damit ist es sinnvoll, das PSA-basierte Alignment, das MODELLER automatisch generiert, durch die MSA-basierte Variante zu ersetzen, wenn Homologiemodelle für die Strukturbibliothek generiert werden sollen.

4.2.6 Modellierungsbeispiel OMPD

Um den Vorteil des MSA-basierten Zuordnungsverfahrens zu veranschaulichen, wurde im Folgenden ein Modellierungsbeispiel gewählt, das besonders deutlich vom MSA

profitiert. Das hier vorgestellte Testprotein ist OMPD. Die Struktur des Targets stammt aus *Methanobacterium thermoautotrophicum* (1lor), die des Templats aus *Saccharomyces cerevisiae* (1qdw).

Die Sequenzidentität von Target- und Templatsequenz liegt insgesamt bei 20%, für das aktive Zentrum beträgt sie 60%. Wenn man die ganze Sequenz betrachtet, werden durch das PSA 23% der Referenzzuordnung gefunden und durch das MSA 49%. Beschränkt man sich auf das aktive Zentrum, so werden durch das PSA 12% und durch das MSA 84% gefunden.

Mit diesen beiden Zuordnungen und der idealen Zuordnung durch das strukturbasierte PSA wurden anschließend Homologiemodelle erzeugt. Der RMSD (aller Atome) zwischen den Modellstrukturen und der Referenz betrug beim PSA 9,2Å für die ganze Struktur und 7,4Å für das aktive Zentrum. Beim MSA waren es jeweils 5,4Å und 4.2 Å. Bei einem idealen Alignment war der RMSD 3.3Å und 2.8Å.

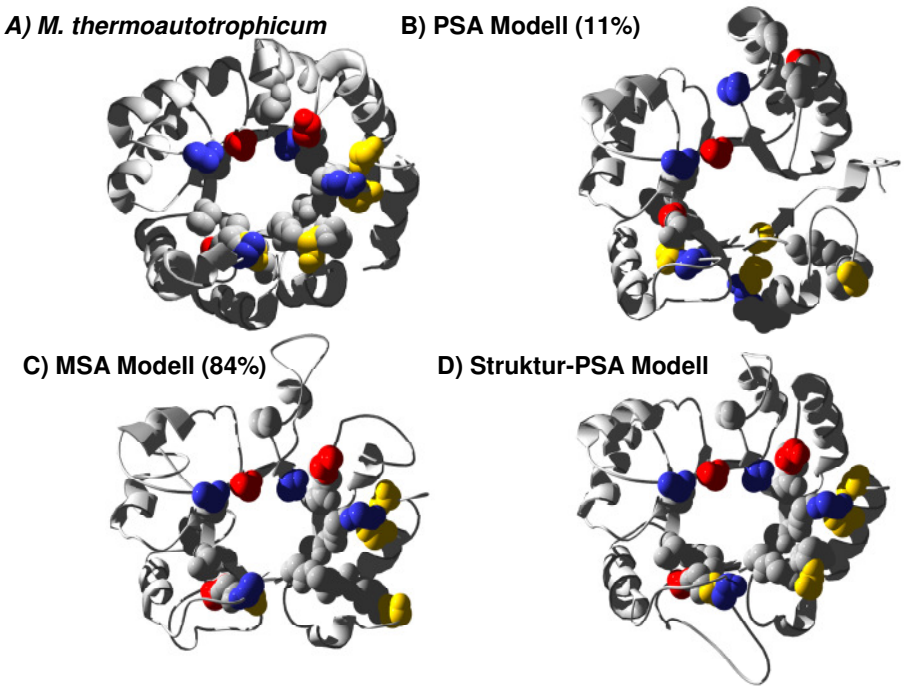


Abbildung 22: OMP Decarboxylase - Zuordnung durch PSA, MSA und SPSA
Hier sind die wildtypische Struktur (*A*) und drei Strukturmodelle der OMP Decarboxylase aus *M. thermoautotrophicum* dargestellt. Die Modelle wurden mit MODELLER erstellt und basieren auf einem PSA (*B*), einem MSA (*C*) oder dem idealen strukturbasierten PSA (*D*). Die Reste der aktiven Zentren sind farblich hervorgehoben. Die Färbung zeigt Rückgratpositionen von Arg, His sowie Lys in blau. Asp und Glu sind rot eingefärbt. Asn, Cys, Gln, Ser, Thr sowie Tyr sind gelb gefärbt. Die hydrophoben Reste sind grau dargestellt. Bei den Modellen *B* und *C* steht in Klammern jeweils der Prozentsatz der Reste, die an richtigen Positionen modelliert wurden.

In diesem Beispiel sind die Vorteile des MSAs sehr deutlich (Abbildung 22). Während durch das PSA die gesamte Struktur und auch das aktive Zentrum falsch modelliert

wurden, nähert sich das Modell auf Basis des MSAs der bestmöglichen Modellierung durch das SPSA an. Setzt man für akzeptable Vorhersagen eine Schwelle von 2Å, ist allerdings auch das MSA in diesem Fall nicht ausreichend, um ein geeignetes Modell zu generieren. Weiterhin wird deutlich, dass die Alignmentqualität zwar großen Einfluss auf die Modellierungsqualität hat, aber nicht für alle Modellierungsprobleme verantwortlich ist. Der RMSD von 3.3Å, der sich aus dem Vergleich der Modellierung mit idealem Alignment und der Referenzstruktur ergibt, zeigt, dass auch die Ähnlichkeit der Vorlagestruktur zur Referenzstruktur und die Leistungsfähigkeit der Modellierungsroutinen für Abweichungen verantwortlich sein müssen.

4.2.7 Sequenzidentität

Im vorhergehenden Kapitel ist als Beispiel die Homologiemodellierung von OMPD vorgestellt worden. Es wurde deutlich, dass ein MSA-basiertes Verfahren die Modellierungsqualität gegenüber einem PSA-basierten Verfahren deutlich erhöht. Trotzdem war die Modellierungsqualität gegenüber dem strukturbasierten Verfahren eher niedrig. Eine Ursache ist die in diesem Fall geringe Sequenzähnlichkeit von nur 20%. Die Autoren von MODELLER (Eswar et al., 2007; Sali & Blundell, 1993) klassifizieren den Bereich mit einer Sequenzähnlichkeit zwischen 20% und 30%, als "*Twilight Zone*" (Rost, 1999). Für diesen Bereich werden Modellierungsprobleme als besonders schwierig eingestuft. Erst ab einer Sequenzähnlichkeit von über 30% klassifizieren die Autoren die Modellierungen als einfach und geben für den zu erwartenden mittleren Fehler einen RMSD-Wert von 3.5Å an. Der verbleibende Bereich, mit einer Sequenzidentität < 10% wird „*Midnight Zone*" genannt. Hier ist die Sequenzähnlichkeit nicht mehr signifikant und es ist überhaupt fraglich, ob die Vorlagestruktur homolog zur modellierenden Struktur ist.

Für die Strukturbibliothek von TRANSCENT ist jedoch nur das aktive Zentrum von Bedeutung. Da in aktiven Zentren die Sequenzidentität allgemein höher ist, ist auch der zu erwartende Modellierungsfehler kleiner. Da ein aktives Zentrum aber in die Struktur eingebettet ist, lässt sich der Modellierungsfehler für die übrige Struktur jedoch nicht einfach ignorieren. In Kapitel 4.2.1 wurde bereits erörtert, dass es fraglich ist, ob die Klassifikation *„Sequenzidentität > 30%"* für das aktive Zentrum allein ein geeignetes Abschätzungskriterium darstellt. Um diese Fragestellung zu untersuchen, wurden mit dem Testdatensatz Proteinmodelle generiert und analysiert (Abbildung 23).

Zuerst wurden alle 62 nicht trivialen Proteinmodelle mit einer idealen Zuordnung modelliert. Dazu wurde ein strukturbasiertes Sequenzalignment verwendet. Die Sequenzidentität variiert in diesen Fällen zwischen 20% und 85%. Der RMSD für die Strukturen bewegt sich zwischen 1.2Å und etwa 4Å. Insgesamt korrelieren Sequenzidentität und RMSD mit einem Korrelationskoeffizient von -0.90. Der mittlere RMSD

liegt bei 2.7Å. Werden ausschließlich Modelle mit einer Sequenzidentität über 30% berücksichtigt, ergibt sich ein mittlerer RMSD von 2.5Å.

Anschließend wurde die Modellierungsqualität für die aktiven Zentren der Modelle bestimmt. Die Sequenzidentität schwankt hier nur zwischen 35% und 100%, was an der stärkeren Konservierung aktiver Zentren liegt. Die RMSD-Werte liegen zwischen 0.5Å und 4Å. Sequenzidentität und RMSD korrelieren mit einem Korrelationskoeffizienten von -0.78. Der durchschnittliche RMSD-Wert liegt bei 2Å. Beschränkt man sich auf Modelle mit mehr als 60% Sequenzidentität, liegt der durchschnittliche RMSD bei 1.6Å und kein Modell hat einen RMSD größer als 3Å.

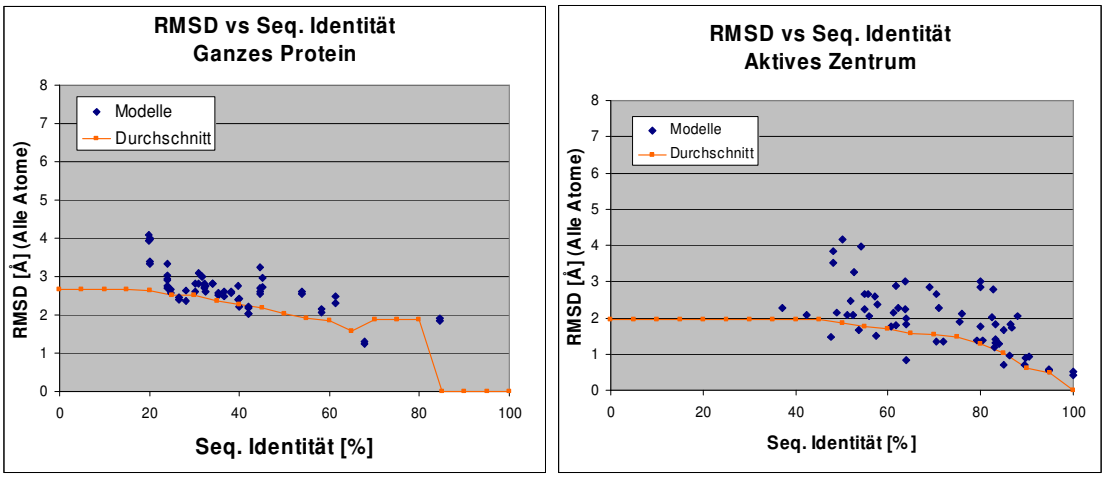


Abbildung 23: Ideales Alignment; Vergleich von Sequenzidentität und RMSD
Im linken Diagramm ist die Sequenzidentität einzelner Modelle (blau) für den Testdatensatz gegen den Modellierungsfehler aufgetragen. Die Modelle wurden auf Basis eines idealen Alignments berechnet. Die Linie (orange) gibt für jeden Sequenzidentitätswert *x* den mittleren RMSD aller Modelle an, deren Sequenzidentität höher als *x* ist. Im rechten Diagramm ist der gleiche Zusammenhang dargestellt, aber auf das aktive Zentrum der Proteinmodelle beschränkt.

Bei dieser Analyse basiert die Modellierung auf einem idealen Alignment. Dieses kann jedoch normalerweise für Homologiemodellierungen nicht generiert werden, da hierfür die Referenzstrukturen fehlen. Trotzdem sind die Ergebnisse wichtig. Da das verwendete Alignment optimal ist, können die hier berechneten Modelle nicht mehr durch ein alternatives Alignment verbessert werden. Bei einer Sequenzidentität von mehr als 60% hat kein aktives Zentrum einen RMSD über 3Å und der durchschnittliche RMSD liegt bei 1.6Å. Wird eine Modellierungsqualität auf diesem Niveau gefordert, so stellt 60% Sequenzidentität eine Untergrenze dar. Das bedeutet, dass die Schwelle für suboptimale Alignments bei mindestens 60% liegen muss, damit eine vergleichbare Modellierungsqualität gewährleistet werden kann.

Im Anwendungsfall wird die Zuordnung aus einem MSA abgeleitet. Obwohl die Sequenzpositionen genauer zugeordnet werden als mit einem PSA, stimmen die

Zuweisungen nicht immer mit der idealen Zuweisung überein. Um eine Aussage für Modellierungen auf Basis eines MSAs machen zu können, wurde die Modellierung der Testproteine unter Verwendung von MSAs wiederholt.

Die Sequenzidentität variiert in diesem Fall zwischen 15% und 85%. Dies ist ein deutlicher Hinweis auf Zuordnungsfehler, denn beim idealen Alignment lag die geringste Sequenzähnlichkeit für den Testdatensatz bei 20%. Die Modelle weichen von den Referenzstrukturen mit RMSDs von 1.5Å bis zu 7Å ab. Fast alle Modelle haben einen größeren RMSD als 3Å. Eine Korrelation zwischen Sequenzidentität und RMSD ist weiterhin messbar, der Korrelationskoeffizient beträgt jedoch nur noch -0.82. Der durchschnittliche RMSD-Wert liegt bei 4.2Å. Bei den Modellen mit einer Sequenzidentität von mindestens 30% liegt der durchschnittliche RMSD-Wert bei 4Å. Damit ist die Modellierungsqualität auf Basis von MSAs deutlich schlechter als auf Basis eines idealen Alignments. Wenn nur Modelle mit einer Sequenzidentität von mindestens 40% berücksichtigt werden, sinkt der mittlere RMSD auf 3Å.

Für die Strukturbibliothek ist allerdings nicht die Qualität des gesamten Modells, sondern vor allem die der aktiven Zentren von Bedeutung. Daher wurde ebenfalls die Modellierungsqualität der aktiven Zentren analysiert (Abbildung 24).

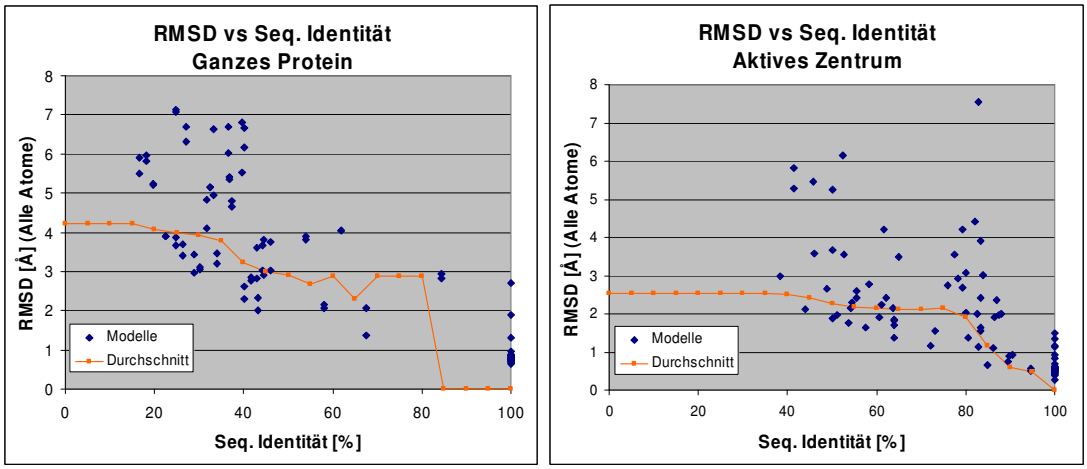


Abbildung 24: MSA basiertes Alignment; Vergleich von Sequenzidentität und RMSD
Im linken Diagramm ist die Sequenzidentität einzelner Modelle (blau) für den Testdatensatz gegen den Modellierungsfehler aufgetragen. Die Modelle wurden auf Alignments berechnet, das aus einem MSA abgeleitet wurde. Die Linie (orange) gibt für jede Sequenzidentität *x* den mittleren RMSD aller Modelle an, deren Sequenzidentität höher als *x* ist. Im rechten Diagramm ist der gleiche Zusammenhang dargestellt, aber auf das aktive Zentrum der Proteinmodelle beschränkt.

In diesem Fall liegt die Sequenzähnlichkeit zwischen 40% und 100%. Der RMSD liegt zwischen 1.5Å und 7.5Å. Der deutliche Unterschied zu den RMSD-Werten bei idealen Alignments ist ein Hinweis auf gravierende Zuordnungsfehler. Die Korrelation zwischen Sequenzidentität und RMSD ist schwächer als mit den idealen Alignments. Der

Korrelationskoeffizient beträgt nur noch -0.65. Der durchschnittliche RMSD-Wert liegt bei 2.5Å. Bei einer Sequenzidentität größer 60% liegt der durchschnittliche RMSD immer noch bei 2.1Å. Dabei ist der RMSD der meisten aktiven Zentren zwar kleiner als 3Å, aber bei einigen Modellen liegt er auch deutlich darüber (bis zu 7.5Å).

Damit lassen sich die aktiven Zentren durch die 60% Schwelle nicht zuverlässig als geeignet klassifizieren. Eine bessere Schwelle anzugeben ist schwierig, da erst bei 80% Sequenzidentität der durchschnittliche RMSD auf 2Å sinkt und selbst dann einige Modelle noch einen RMSD von über 4Å aufweisen.

Trotz der Verwendung von MSAs entstehen also Zuordnungsfehler, welche die Modellierungsqualität deutlich herabsetzen. Die Ergebnisse lassen vermuten, dass eine zuverlässige Abschätzung der Modellierungsqualität auf Basis der Sequenzidentität allein nicht möglich ist.

4.2.8 MSA-Alignmentqualität

Im vorhergehenden Kapitel wurde gezeigt, dass für MSA-basierte Modelle der Sequenzidentitätswert alleine keine ausreichend genaue Abschätzung der Modellierungsqualität erlaubt. Das Problem sind die Zuordnungsfehler im Alignment, das der Modellierung zu Grunde liegt. Das Problem liegt also in der Qualität der MSAs, aus denen die Alignments für die Modelle abgeleitet worden sind. Das hier verwendete MSA-Programm MAFFT (Katoh et al., 2002) ist zur Zeit einer der besten Ansätze (Wallace et al., 2006). Nicht notwendigerweise ist das MSA-Programm selbst die Ursache für die schlechte Qualität des MSAs. Das Problem kann auch in den Eingabedaten liegen. Wenn die Sequenzen, aus denen das MSA erzeugt werden soll, nicht ähnlich genug sind, entstehen auch mit performanten MSA-Programmen Zuordnungsfehler. Daher wird hier nicht versucht, die Modellierungsqualität durch Verbesserung der MSA-Alignmentqualität zu erhöhen. Stattdessen wird eine bestimmte MSA-Alignmentqualität gefordert, um die Modellierungsqualität zu gewährleisten. Die Alignmentqualität wird hier mit Hilfe der T-COFFEE Routine T-COFFEE-CORE (Notredame et al., 2000) bestimmt (Abbildung 25).

Für jedes Modell wurden die CORE-Werte berechnet. Diese liegen zwischen 1% und 75%. Der Korrelationskoeffizient für CORE-Wert und RMSD ist -0.47. Werden die 38 Modelle mit CORE-Werten über 20% separat betrachtet, ergibt sich ein mittlerer RMSD von 3.7Å, für die 24 übrigen Modelle ist der mittlere RMSD-Wert 5.1Å.

Für die aktiven Zentren wurden die CORE-Werte separat bestimmt. Hier schwanken die Werte zwischen 1% und 98%. Der Korrelationskoeffizient für CORE-Werte und RMSD-Werte beträgt -0.52. Werden hier die 38 Modelle mit den höchsten CORE-Werten separat

betrachtet, ergibt sich ein mittlerer RMSD von 2Å. Für die 24 übrigen Modelle ist der mittlere Wert 3.5Å.

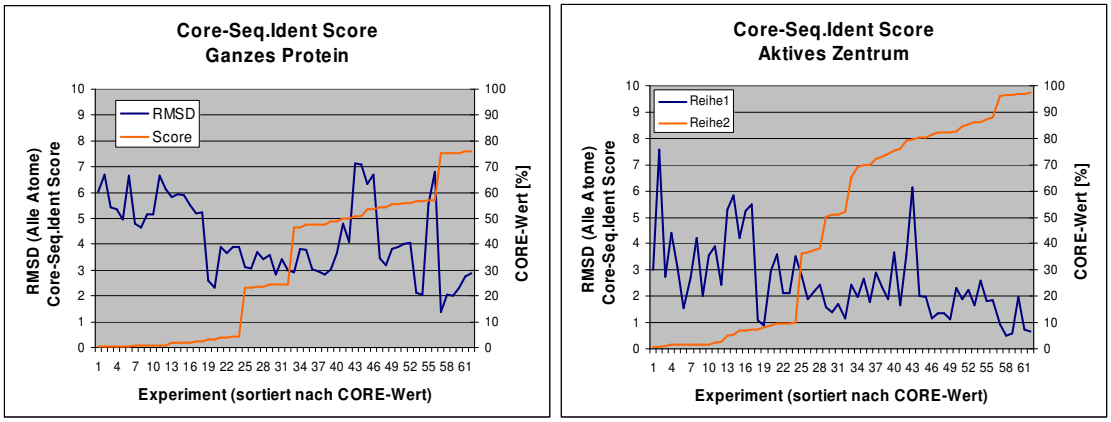


Abbildung 25: Abhängigkeit von MSA-Qualität und Modellierungsgenauigkeit
Im linken Diagram sind zwei Kurven überlagert. Diese sind die Modellierungsgenauigkeit als RMSD für Modelle der einzelnen Testproteine (blau) und die MSA-Qualität als CORE-Wert (orange). Die Testproteine sind nach dem CORE-Wert sortiert. Im rechten Diagram sind die gleichen Zusammenhänge für die aktiven Zentren der Modelle geplottet.

Werden nun Schwellen für eine minimale Sequenzidentität und Schwellen für eine minimale MSA-Alignmentqualität kombiniert, so lassen sich die Modelle identifizieren, die beim MSA-basierten Alignment schlechte Voraussetzungen mitbringen (Abbildung 26).

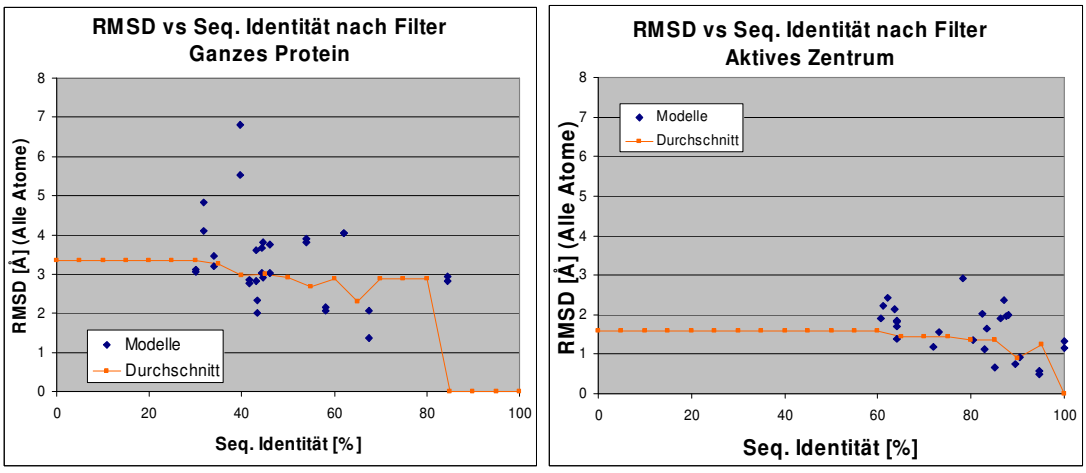


Abbildung 26: RMSD bei ausreichender Sequenzidentität und MSA-Qualität
Im linken Diagramm ist die Sequenzidentität einzelner Modelle (blau) für den Testdatensatz gegen den Modellierungsfehler aufgetragen. Die Modelle wurden aus Alignments berechnet, die aus einem MSA abgeleitet wurden. Die Linie (orange) gibt für jede Sequenzidentität *x* den mittleren RMSD aller Modelle an, deren Sequenzidentität höher als *x* ist. Alle Modellierungen erfüllen die Schwellen Sequenzidentität > 40% und CORE-Wert > 20%. Im rechten Diagramm ist der gleiche Zusammenhang dargestellt, aber auf das aktive Zentrum der Proteinmodelle beschränkt. Hier sind die Schwellen Sequenzidentität > 60% und CORE-Wert > 20%.

Liegt die Schwelle für die minimale Sequenzidentität bei 40% und die Schwelle für die minimale MSA-Qualität bei 20%, so bleiben 22 von 62 Modellen übrig. Der RMSD Modelle oberhalb beider Schwellen liegt bei höchstens 4Å, der Mittelwert bei 3Å.

Wird für die aktiven Zentren ein Schwellwert von mindestens 60% Sequenzidentität und mindestens 20% MSA-Alignmentqualität angewendet, bleiben 26 von 62 Modellen übrig. Der RMSD-Wert liegt dann immer unter 3Å und im Mittel bei 1.6Å.

Die beiden Schwellen „*Sequenzidentität > 40%*“ und „*CORE-Wert > 20%*“ sind bei der Generierung aller Strukturbibliotheken verwendet worden, die im Rahmen der vorliegenden Arbeit benötigt wurden.

4.3 Gewichte und Performanz

Im Programm TRANSCENT werden die Energieterme der vier Module in einer Energiefunktion miteinander kombiniert. Auf diese Weise können alle Module gleichzeitig den Optimierungsprozess beeinflussen. Damit der Einfluss keines Moduls überwiegt, müssen die Energieterme durch eine geeignete Gewichtung aufeinander abgestimmt werden. Um eine optimale Gewichtung zu finden, werden im Folgenden Testdesigns unter Verwendung von Proteinen aus einem Testdatensatz durchgeführt.

4.3.1 Evaluation mit Testdesigns

Um die einzelnen Module von TRANSCENT aufeinander abzustimmen und um die Leistungsfähigkeit für Modulkombinationen zu testen, werden in dieser Arbeit Testdesigns verwendet. Bei einem Testdesign wird das aktive Zentrum eines Proteins auf sich selbst „*transferiert*“. Für ein Programm mit guter Performanz sollte dieser Testfall besonders einfach zu modellieren sein, da die Struktur optimal zum geforderten aktiven Zentrum passt. Das resultierende Proteinmodell wäre dann der wildtypischen Struktur sehr ähnlich. Auf diese Weise wurde bereits die Leistung von verschiedenen ROSETTA Programmen (Kuhlman & Baker, 2000; Meiler & Baker, 2006; Zanghellini et al., 2006) und auch von DEZYMER (Hellinga & Richards, 1991) gemessen.

Um ein Testdesign mit einem Protein durchzuführen, wird für TRANSCENT eine Modulkombination festgelegt und es werden vom Protein die Seitenketten des aktiven Zentrums entfernt. Somit sind alle Reste des aktiven Zentrums Glycine, d.h. das Testprotein hat jetzt ein „leeres“ aktives Zentrum. Anschließend wird mit TRANSCENT das aktive Zentrum neu modelliert, indem es optimale neue Seitenketten für diese Positionen sucht. Das resultierende Modell wird dann mit der ursprünglichen Proteinstruktur verglichen (Abbildung 27). Die Ähnlichkeit der Modelle zu den Strukturen, von denen ausgegangen wurde, ist ein Maß für die Leistung der Modulkombination. Um die Leistungsfähigkeit zuverlässig beurteilen zu können, werden solche Testdesigns für einen größeren Testdatensatz von Proteinen durchgeführt.

Damit die Ähnlichkeit quantifiziert werden kann, werden zwei verschiedene Maße verwendet. Das erste Maß ist die Sequenzidentität. Hierfür wird im aktiven Zentrum der Anteil der Positionen bestimmt, bei denen die Aminosäuren des Modells mit dem Wildtyp übereinstimmen. Die Sequenzidentität ist ein sehr anschauliches Maß, liefert aber keine Informationen über die Ähnlichkeit der nicht identisch modellierten Aminosäuren. Daher wird als zweites Maß der durchschnittliche BLOSUM-Score für die Positionen des aktiven Zentrums ermittelt. Der BLOSUM-Score erlaubt auch die nicht

identisch modellierten Positionen zu beurteilen. Damit ist er ein präziseres, aber weniger intuitiv interpretierbares Maß für die Ähnlichkeit.

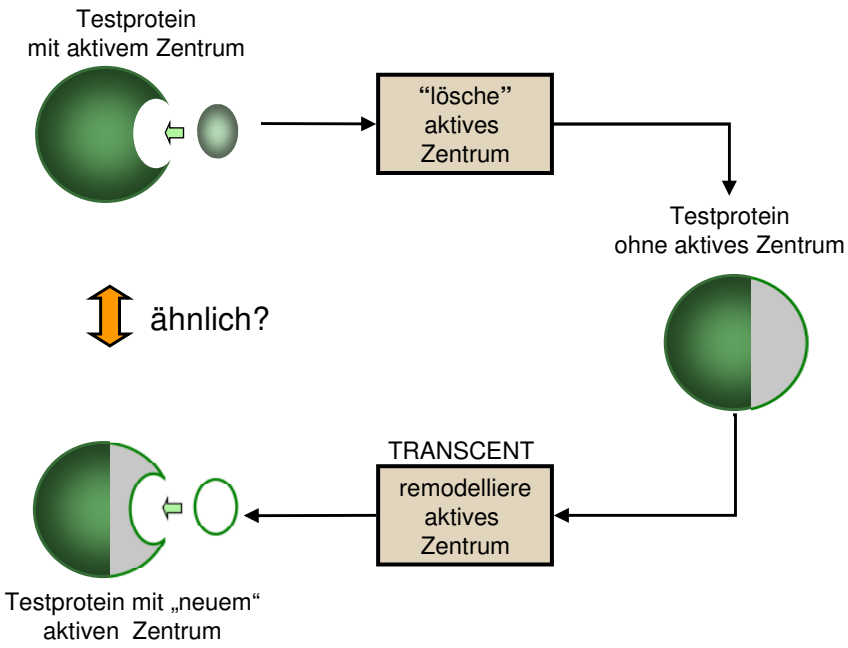


Abbildung 27: Flussdiagramm für ein Testdesign
Die Abbildung zeigt den Ablauf eines Testdesigns, mit dem die Leistungsfähigkeit der Module von TRANSCENT evaluiert wird. Dazu wird das bekannte aktive Zentrum eines Testproteins „gelöscht“, indem alle Seitenketten als Glycine modelliert werden. Anschließend wird TRANSCENT verwendet, um das gelöschte aktive Zentrum neu zu modellieren. Die Ähnlichkeit zwischen dem generierten Modell und dem wildtypischen Testprotein ist ein Indikator für die Leistung.

Die vollständige Optimierung der Proteinsequenz ist rechenaufwändig. Da mit den Testdesigns die Güte von modellierten aktiven Zentren erfasst werden soll, kann sich die Optimierung auf die Region des aktiven Zentrums beschränken. Auf diese Weise lässt sich der Rechenaufwand für die Optimierung reduzieren. Die Strukturen der einzelnen Proteine werden dazu in drei Regionen partitioniert (Abbildung 28).

Die erste Region umfasst alle Positionen des aktiven Zentrums. Als aktives Zentrum werden die Positionen klassifiziert, die höchstens 7Å vom Liganden entfernt sind. In dieser Region sind die Aminosäuren frei wählbar. Dafür stehen alle Rotamere der Rotamerbibliothek zur Verfügung. In der zweiten Region sind nur die Seitenketten flexibel. Diese Region besteht aus Positionen, die zwischen 7Å und 15Å vom Liganden entfernt liegen. Die Rotamere dieser Positionen sind auf die wildtypischen Aminosäuren beschränkt. Dadurch bleibt die Region um das aktive Zentrum flexibel und ein neu modelliertes aktives Zentrum kann spannungsfreier in die Struktur eingebettet werden. Die dritte Region umfasst alle verbleibenden Positionen des Proteins. Für diese Posi-

tionen werden die Reste aus der wildtypischen Struktur übernommen und starr gehalten. Es wird davon ausgegangen, dass die Reste der dritten Region keinen Einfluss auf die Positionen des aktiven Zentrums haben. Damit ein Protein einfach in diese drei Regionen unterteilt werden kann, wird eine Struktur mit einem im aktiven Zentrum gebundenen Liganden benötigt.

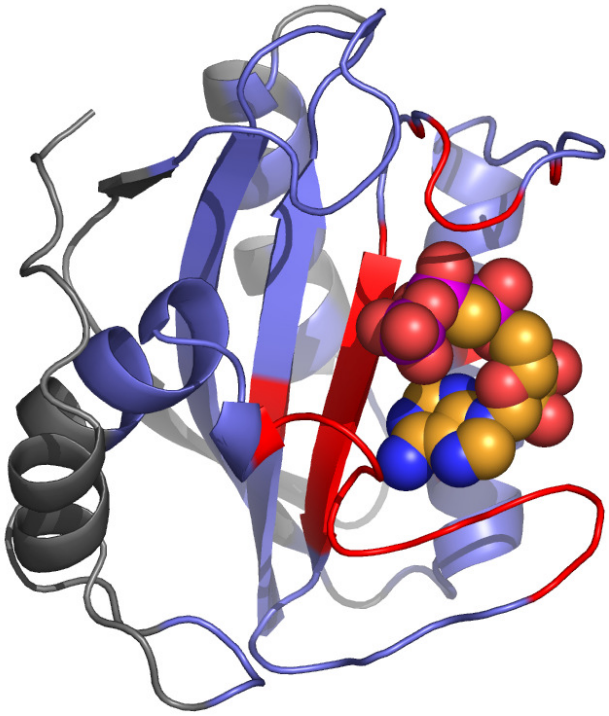


Abbildung 28: Aufteilung der Positionen einer Proteinstruktur für ein Testdesign
Die Positionen der abgebildeten Proteinstruktur (1f9y) werden für ein Testdesign in drei Regionen aufgeteilt. Als Region des aktiven Zentrums (rot) sind alle Positionen definiert, deren Reste einen maximalen Abstand von 7Å zum Liganden (Kugeldarstellung) haben. Hier sind für das Design alle Aminsäuren frei wählbar. Die zweite Region (blau) besteht aus den Positionen, deren Reste einen Abstand zwischen 7Å und 15Å zum Liganden haben. Hier sind nur die Rotamere der wildtypischen Aminosäuren erlaubt. Die dritte Region (grau) besteht aus den verbleibenden Positionen, deren Reste aus der Kristallstruktur als starr übernommen werden.

4.3.2 Testdatensatz

Der Testdatensatz, mit dem die Testdesigns zur Leistungsbestimmung von TRANSCENT durchgeführt wurden, umfasst 128 Proteine mit gebundenem Liganden. In Tabelle 5 sind die PDB-Codes der Proteine aufgeführt.

Tabelle 5: PDB-Codes der Proteine im Testdatensatz

PDB-Code									
1a4i	1a6m	1ajs	1b8o	1ccw	1d2s	1d4o	1dbt	1dqx	1ds1
1dzk	1eix	1f74	1f8e	1f9v	1f9y	1fp2	1fs7	1ft5	1g3p
1g6s	1ghe	1gwe	1h4g	1hqs	1hx0	1jcm	1jfb	1jub	1jx6
1k3y	1kly	1km4	1kqp	1kt6	1lbf	1lbm	1ltz	1lyc	1m0k
1m15	1m40	1me4	1n08	1n8k	1nox	1o08	1o8b	1obd	1obo
1ox5	1po5	1q0r	1q6z	1q92	1qnr	1qop	1qwo	1qxy	1r2q
1r5l	1rcq	1rp0	1rwh	1rya	1s1d	1sg4	1si6	1su8	1t2d
1tbf	1tjy	1tt8	1u0f	1u4b	1u7g	1uas	1ucd	1ujp	1usc
1v2x	1vk5	1vyr	1w0h	1w0p	1w66	1w6g	1wb4	1wbe	1wdd
1wqw	1wui	1wvf	1wvq	1x7d	1x8q	1x9i	1xdn	1xg0	1xg4
1y0y	1y3n	1ymt	1z2n	1z53	1z6f	1zdy	1zhx	1zk4	1zr6
1zz1	2a50	2a84	2aeb	2apj	2asb	2b82	2bfd	2bkx	2bln
2bog	2bzg	2c1v	2czl	2f5t	2f6u	2nlr	2tys		

Die Proteine wurden mit folgenden Filterkriterien aus der PDB-Datenbank (Bernstein et al., 1977) ausgewählt:

1. Es wurden nur Strukturen mit einer Auflösung von höchstens 1.6Å und einem R-Faktor von höchstens 0.25 verwendet. Im paarweisen Vergleich sind die Sequenzen der Proteinstrukturen maximal 20% identisch. Bestimmt wurden diese Strukturen mit dem Culling-Server der Dunbrack Gruppe (Wang & Dunbrack, 2003; Wang & Dunbrack, 2005)
2. Es wurden nur Strukturen verwendet, die durch Röntgenkristallographie aufgeklärt worden sind und aus mindestens 100 Aminsäuren bestehen.
3. In jeder Struktur muss genau ein Ligand aus mindestens 10 Atomen gebunden sein.
4. Falls die PDB-Datei mehrere Proteinketten beinhaltet, wurde nur die Kette verwendet, die dem Liganden am nächsten war.
5. Der Ligand musste von wenigstens 10 Resten in einem Abstand von höchstens 5Å umgeben sein.

Durch die Filterkriterien wird sichergestellt, dass die Proteinstrukturen hohe Qualität haben und nicht zu ähnlich zueinander sind. Jedes Protein hat zudem ein aktives Zentrum mit gebundenem Liganden. Beispiele für die Testproteine sind in Abbildung 29 dargestellt.

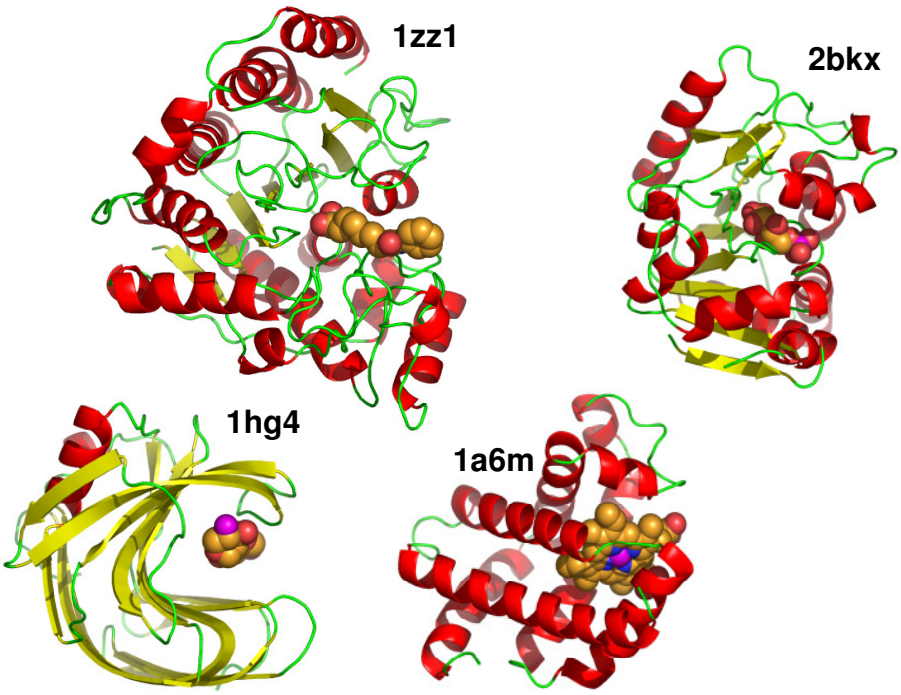


Abbildung 29: Proteinstrukturen aus dem Testdatensatz
In der Abbildung sind vier Proteine des Testdatensatzes in Bänder-Darstellung zu sehen. Jedes der Proteine hat einen Liganden (Kugeldarstellung) gebunden. Zu den Proteinen ist der PDB-Code angegeben.

4.3.3 Vergleich: ROSETTA und EGAD

Das Basismodul in TRANSCENT ist das Proteindesign-Modul zur Bewertung der Stabilität. Es beruht auf Rotamer- und Energietabellen, die von externen Programmen zur Verfügung gestellt werden. Dafür stehen die quelloffenen Programme ROSETTA DESIGN und EGAD zur Auswahl. Mit dem Testdatensatz wurde die Leistungsfähigkeit der beiden Programme verglichen.

Tabelle 6: Fünf Proteine für den Leistungsvergleich von ROSETTA DESIGN und EGAD
Die Tabelle beschreiben für jedes Protein im Einzelnen: den PDB-Code, den Proteinnamen, die Anzahl der variablen Positionen im aktiven Zentrum sowie die Größe der Energietabellen, die ROSETTA DESIGN und EGAD generiert haben.

PDB-Code	Protein	Pos.im aktiven Zentrum	Energietabelle ROSETTA / EGAD
1qwo	Phytase	11	20 MB / 136 MB
1vk5	Protein mit unbekannter Funktion	12	27 MB / 174 MB
1xg0	Phycoerythrin 545	18	19 MB / 306 MB
1rya	GDP-Mannose Mannosyl Hydrolase	20	34 MB / 384 MB
1d4o	Domäne III der Transhydrogenase	27	84 MB / 537 MB

Bereits beim Generieren der Energietabellen für die Testproteine wurde ein erster Unterschied deutlich. Während im Fall von ROSETTA DESIGN der Speicherbedarf für keine Energietabelle 500 MB überschritt, konnten für EGAD viele Energietabellen nicht generiert werden, weil der Hauptspeicher (2 GB) der verwendeten Computer nicht ausreichte. Um trotzdem einen Eindruck von der Leistungsfähigkeit der beiden Programme zu bekommen, wurden fünf Testproteine (Tabelle 6) ausgewählt, für die sich die Energietabellen auch von EGAD generieren ließen. Die Testdesigns wurden 10-mal wiederholt und die Ergebnisse gemittelt.

Die Werte für die Sequenzidentität der aktiven Zentren liegen bei ROSETTA DESIGN zwischen 9% und 55%. Der Durchschnitt liegt bei 35%. Der BLOSUM-Score liegt zwischen 0.6 und 3.4 und im Mittel bei 1.7. Die Sequenzidentität korreliert mit dem BLOSUM-Score mit einem Korrelationskoeffizienten von 0.90. Bei EGAD liegt die Sequenzidentität zwischen 13% und 33% und im Mittel bei 18.5%. Der BLOSUM-Score liegt zwischen -0.7 und 0.7. Der Mittelwert beträgt 0.1. Bei EGAD korreliert die Sequenzidentität und der BLOSUM-Score mit einem Korrelationskoeffizienten von 0.57.

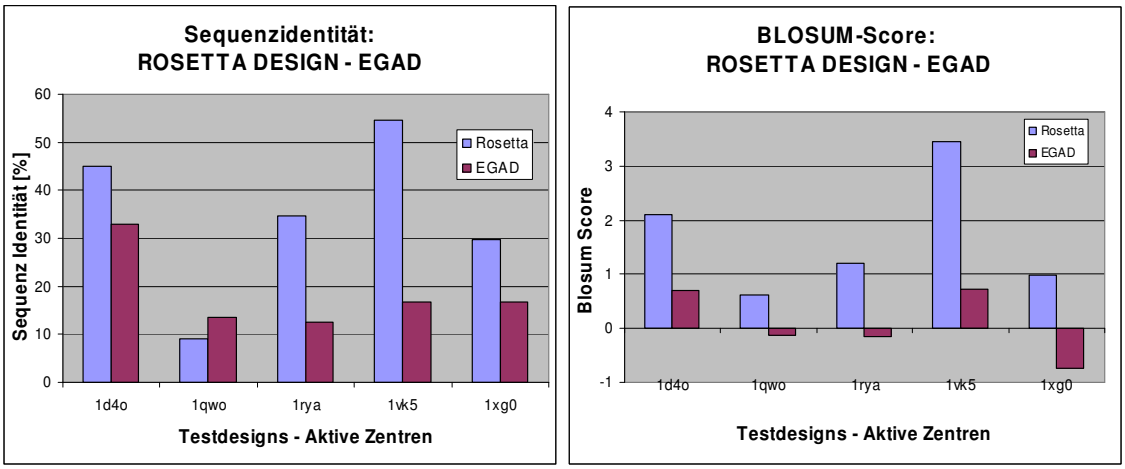


Abbildung 30: Vergleich von EGAD und ROSETTA DESIGN - fünf Testproteine
Für die fünf Testproteine wurden Testdesigns mit ROSETTA (blau) und EGAD (lila) durchgeführt und 10-mal wiederholt. Im linken Diagramm ist die mittlere Sequenzidentität der Modelle aufgeführt, im rechten der entsprechende BLOSUM-Score.

Im direkten Vergleich der einzelnen Testdesigns war die Sequenzidentität für die Modelle auf Basis von ROSETTA DESIGN in vier von fünf Fällen besser als EGAD, bezogen auf den BLOSUM-Score sogar in allen Fällen. Sequenzidentität und BLOSUM-Score korrelieren bei ROSETTA DESIGN viel stärker als bei EGAD. Dies legt den Schluss nahe, dass die Sequenzidentität die Leistung von EGAD nicht geeignet beschreibt. Die Werte deuten Probleme bei den Positionen an, die nicht identisch modelliert worden sind.

Um dies näher zu untersuchen, wurde die Aminosäureverteilung für die Positionen der aktiven Zentren ausgewertet. Dabei wurden für jede der 20 Aminosäuren drei Häufigkeiten bestimmt: 1) Die Häufigkeit mit der die Aminosäure im aktiven Zentrum der fünf wildtypischen Proteine zu finden ist. 2) Die Häufigkeit mit der die Aminosäure richtig in den Testdesigns modelliert worden ist. 3) Die Häufigkeit mit der die Aminosäure an einer falschen Stelle verwendet worden ist (Abbildung 31).

Obwohl die Datenbasis von nur fünf Testproteinen keine detaillierte Analyse erlaubt, werden problematische Aminosäure-Präferenzen im Fall von EGAD offensichtlich. In 29 von 98 Fällen ist Tryptophan als Aminosäure gewählt worden. Zwar wurden die beiden wildtypischen Tryptophane wieder gefunden, aber in 27 Positionen kommt Tryptophan in der wildtypischen Sequenz nicht vor. Auch Methionin wird bei EGAD auffällig oft (15 von 98 Fällen) gewählt. Insgesamt besteht eine Tendenz für die Wahl hydrophober Aminosäuren. Die geladenen Aminosäuren Glutamat, Aspartat und Arginin sowie die polaren Aminosäuren Asparagin, Cystein, Glutamin Serin und Threonin wurden gar nicht ausgewählt. ROSETTA DESIGN zeigt dagegen keine derartig auffällige Tendenz zur Überrepräsentation weniger Aminosäuren.

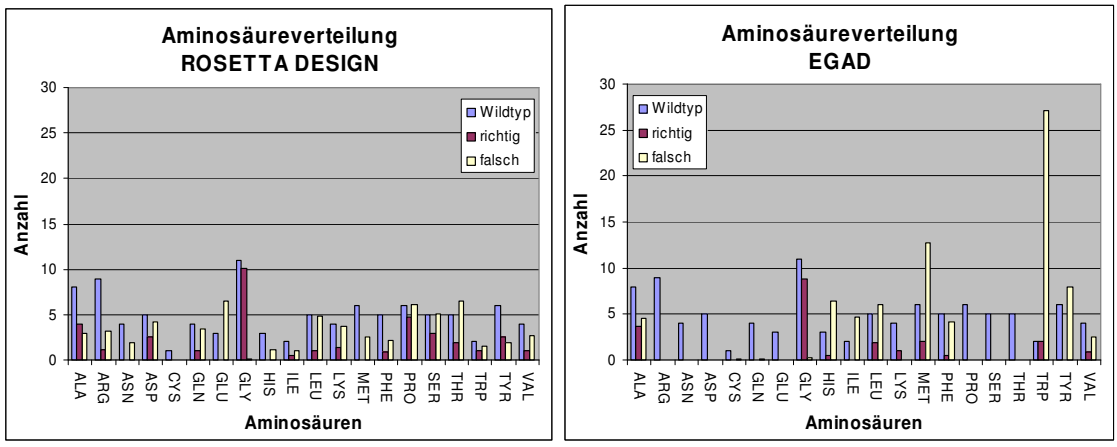


Abbildung 31: Vergleich von EGAD und ROSETTA DESIGN - Aminosäurehäufigkeiten
Um die Leistung von ROSETTA DESIGN (links) und EGAD (rechts) im Detail vergleichen zu können, wurden die Ergebnisse der Testdesigns pro Aminosäure analysiert. Dazu sind für jede Aminosäure deren Vorkommen im aktiven Zentrum der wildtypischen Proteine (blau), sowie die Anzahl richtig gewählter (lila) und falsch gewählter Fälle (gelb) angegeben.

Die Ergebnisse für die fünf Testdesigns lassen vermuten, dass ROSETTA DESIGN besser geeignet ist die Aminosäuren von aktiven Zentren zu modellieren als EGAD. Für eine quantitative Beurteilung ist die Datenbasis jedoch zu gering. Da EGAD aufgrund des höheren Speicherbedarfs auch aus technischer Sicht schwierig zu nutzen ist, beschränken sich alle folgenden Experimente auf die Verwendung von ROSETTA DESIGN.

4.3.4 ROSETTA DESIGN

Um die Leistung des ROSETTA-Moduls zu quantifizieren, wurden Testdesigns für den gesamten Testdatensatz zunächst mit diesem Modul allein ausgewertet. Die Tests wurden 10-mal wiederholt und die Ergebnisse gemittelt.

Die aktiven Zentren der Testdesigns stimmen mit den wildtypischen Proteinsequenzen zwischen 9% und 56% überein. Im Mittel liegt der Wert für die Sequenzidentität bei 31%. Der BLOSUM-Score liegt zwischen -0.4 und 3.4 und im Mittel bei 1.1. Die Sequenzidentitätswerte und die BLOSUM-Scores korrelieren mit einem Korrelationskoeffizienten von 0.88. Die einzelnen Testproteine haben unterschiedlich große aktive Zentren. Die Anzahl der variablen Positionen liegt zwischen 10 und 60. Im Mittel sind es 28 Positionen. Die Größe der aktiven Zentren und die Sequenzidentitätswerte der Testdesigns korrelieren mit einem Korrelationskoeffizienten von 0.23.

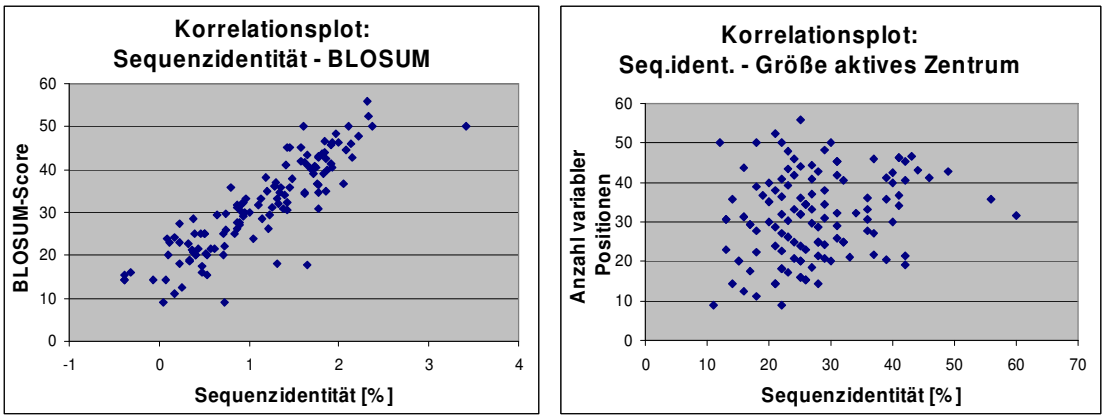


Abbildung 32: ROSETTA DESIGN - Korrelationsplots
Im linken Diagramm ist für 128 Testdesigns der Sequenzidentitätswert gegen den BLOSUM Score aufgetragen. Im rechten Diagramm ist für die Testdesigns der Sequenzidentitätswert gegen die Größe des jeweiligen aktiven Zentrums aufgetragen.

Um die Leistung von ROSETTA DESIGN besser einordnen zu können, sind die publizierten Ergebnisse der Evaluierung des Programms hilfreich (Kuhlman & Baker, 2000). Die Evaluierung wurde mit Hilfe von Testdesigns durchgeführt. Der dabei verwendete Testdatensatz umfasst 108 Proteine. Für die Oberfläche der Testproteine beträgt die Sequenzidentität mit den Testdesigns 27% und für das Innere der Proteine 51%. Da aktive Zentren sich weder eindeutig der Proteinoberfläche noch dem Proteininneren zuordnen lassen, sind die mit dem hier verwendeten Testdatensatz erreichten 31% plausibel.

Die Leistung schwankt sehr stark zwischen den einzelnen Testproteinen, was durch die heterogene Zusammensetzung des Testdatensatzes und die damit verbundenen

Funktionsvielfalt der aktiven Zentren erklärt werden kann. Der niedrige Korrelationskoeffizient zwischen den Sequenzidentitätswerten und der Größe der aktiven Zentren legt nahe, dass die große Streuung der Leistung nicht auf die unterschiedlichen Größen der aktiven Zentren zurückzuführen ist.

Der BLOSUM-Score hingegen korreliert gut mit der Sequenzidentität. Damit sind die Sequenzidentitätswerte repräsentativ für die Gesamtähnlichkeit zwischen den aktiven Zentren der Testproteine und den Modellen. Der erreichte Sequenzähnlichkeitswert ist also auch ein Indikator für die Ähnlichkeit der nicht identisch modellierten Reste.

Abschließend wurde untersucht, ob ROSETTA DESIGN auffällige Präferenzen für einzelne Aminosäuren hat. Dazu wurde die Verteilung jeder Aminosäure ausgewertet. Insgesamt sind 3498 Positionen neu modelliert worden (Abbildung 33). Die Aminosäurekomposition der Modelle korreliert mit der wildtypischen Komposition mit einem Korrelationskoeffizient von 0.90. Die Aminosäuren werden mit einer ähnlichen Häufigkeit für die Modellierung verwendet, wie sie auch in den wildtypischen aktiven Zentren vorkommen. Das ist bemerkenswert, da nur 31% der Aminosäuren übereinstimmend modelliert worden sind.

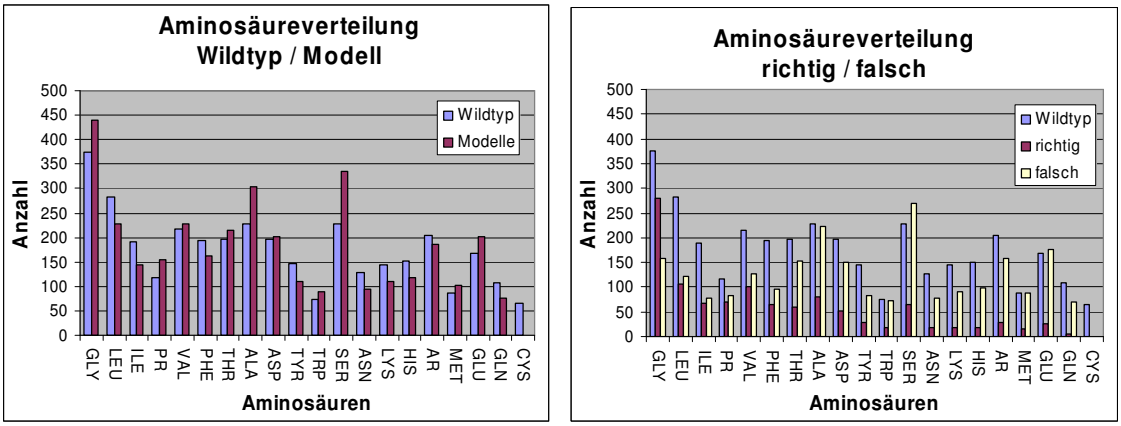


Abbildung 33: ROSETTA DESIGN - Aminosäureverteilungen
Um die Leistung von ROSETTA DESIGN genauer beurteilen zu können, wurde die Aminosäureverteilung für die aktiven Zentren der 128 Testproteine ausgewertet (3498 verschiedene Positionen). Im linken Diagramm ist pro Aminosäuretyp die Häufigkeit im Wildtyp (blau) sowie die Häufigkeit in den Testdesigns (lila) aufgetragen. Im rechten Diagramm ist für jede Aminosäure das Vorkommen im Wildtyp (blau), das mit dem Wildtyp übereinstimmende Vorkommen (lila) sowie das nicht übereinstimmende Vorkommen (gelb) aufgetragen.

Anschließend wurde für jede Aminosäure ausgewertet, wie häufig eine im Modell gewählte Aminosäure dem Wildtyp entspricht und wie oft sie davon abweicht. Glycin schneidet dabei am besten ab. Zum einen wird es in 280 von 375 Fällen gefunden, zum anderen ist nur in 159 Fällen ein Glycin an eine Position modelliert worden, an der im wildtypischen Protein kein Glycin ist. Da Glycin keine Seitenkette hat, weisen Proteinstrukturen an Glycin-Positionen oft Φ-/Ψ-Winkel auf, die mit keiner anderen

Aminosäure erreicht werden können. Der Umstand, dass ROSETTA DESIGN Glycine besonders gut vorhersagt, lässt sich durch die Verwendung einer Rückgrat-abhängigen Rotamerbibliothek erklären.

Cysteine sind dagegen schwierig zu modellierende Aminosäuren, da sie von ROSETTA DESIGN gar nicht verwendet wurden, obwohl sie an 65 Positionen vorkommen. Cysteine sind in aktiven Zentren oft katalytisch essentiell. Im restlichen Protein sind sie aufgrund der Fähigkeit Cystin-Brücken (zwei Cysteine die über die Thiolgruppe kovalent verbunden sind) auszubilden für die Stabilität verantwortlich. Da ROSETTA die Proteinstabilität optimiert, wird es einzelne Cysteine unterdrücken.

Werden die Aminosäuren aufgrund des Verhältnisses von korrekt zu falsch modellierten Positionen sortiert, so sind die Aminosäuren mit dem besten Verhältnis hydrophob (Gly, Leu, Pro, Ile, Val, Phe, Trp, Ala). Diese Tendenz ist jedoch nicht ebenso stark ausgeprägt, wie bei der Verwendung von EGAD (vgl. letztes Kapitel). Zu einem gewissen Grad lässt sich diese Beobachtung damit erklären, dass auch in aktiven Zentren hydrophobe Aminosäuren eher vergraben vorliegen und dass durch diese Einbettung die Wahl der richtigen Aminosäure einfacher wird. Am deutlichsten wird der Effekt für den oft hydrophoben Kern von Proteinen. Da aktive Zentren modelliert werden, ist ein Teil der zu modellierenden Reste katalytisch relevant. Diese Reste sind oft hydrophil. Es konnte schon häufig gezeigt werden, dass die Ersetzung von katalytisch essentiellen Resten Proteine stabilisiert (Nagatani et al., 2007; Shoichet et al., 1995). Da ROSETTA die Stabilität optimiert, werden destabilisierende katalytische Reste nicht bevorzugt.

ROSETTA DESIGN verwendet zwar die verschieden Aminosäuren in einem sehr ausgewogenen Verhältnis, hat aber vor allem Probleme, hydrophile Aminosäuren an die richtigen Positionen zu modellieren. Im Schnitt modelliert das Programm für aktive Zentren 32% der Aminosäuren in Übereinstimmung mit dem Wildtyp.

4.3.5 DRUGSCORE

Das zweite Modul von TRANSCENT verwendet DRUGSCORE, um die Wechselwirkungen zum Liganden im aktiven Zentrum zu beschreiben. Die Wechselwirkungen werden in DRUGSCORE anders berechnet als in ROSETTA DESIGN. Daher müssen die Energien, die mit DRUGSCORE berechnet worden sind, geeignet gewichtet werden, ehe sie zu den ROSETTA-Energien addiert werden können. Um das optimale Gewichtungsverhältnis zu bestimmen, wurde der Testdatensatz mit verschiedenen Gewichten ausgewertet. Gesucht wurde dasjenige Gewicht, bei dem die Auswertung die besten Ergebnisse liefert. DRUGSCORE-Energien nehmen größere Werte an als Energien von ROSETTA DESIGN. Da die Energien in der gleichen Größenordnung liegen müssen, wurde zur Gewichtsbestimmung das Intervall 1/50000 bis

100/50000 durchsucht. Um die Anzahl der Schritte zu beschränken, wurde zwischen den Gewichten die Schrittweite immer um 1.3 angehoben. Auf diese Weise liegen 17 verschiede Gewichte in dem zu untersuchenden Intervall. Die Testdesigns wurden 10-mal wiederholt und die Ergebnisse gemittelt.

Werden die einzelnen Gewichte gegen die erreichte Sequenzidentität aufgetragen, so beschreiben die Punkte eine Kurve, die bei 32% für ein Gewicht von 0 beginnt und bei 25% bei einem Gewicht von 100/50000 endet. Das Maximum der Kurve liegt bei einem Gewicht von 7.5/50000 und beträgt 37%. Die Wahl dieses Maximums als Gewicht ist unkritisch, da auch die benachbarten Gewichte einen ähnlichen Sequenzidentitätswert bewirken. Wenn anstatt der Sequenzidentität der BLOSUM-Score verwendet wird, so ergibt sich ebenfalls für das Gewicht von 7.5/50000 das beste Ergebnis (BLOSUM-Score: 1.5). Insgesamt sind sich die Kurvenverläufe von BLOSUM-Scores und Sequenzidentität sehr ähnlich (Korrelationskoeffizient: 0.997). Daher wird für alle weiteren Experimente der Gewichtsfaktor für DRUGSCORE auf 7.5/50000 festgelegt.

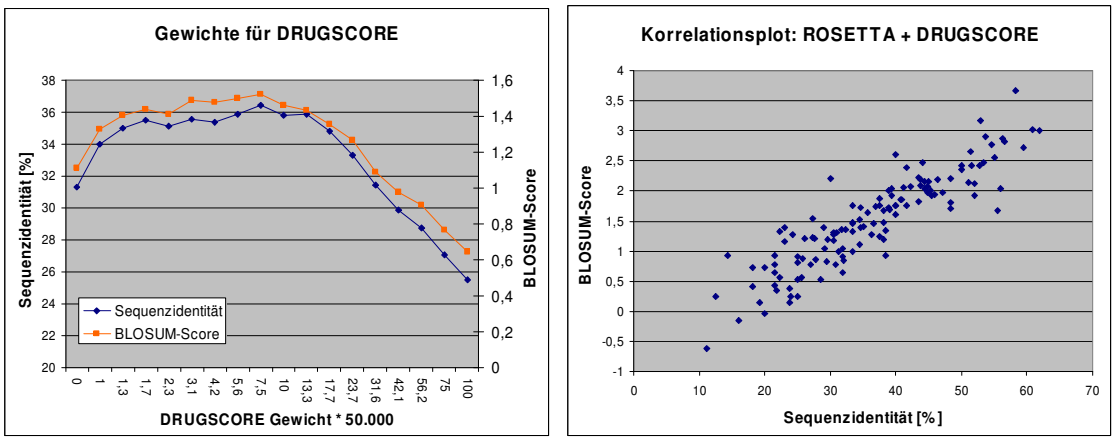


Abbildung 34: DRUGSCORE Modul – Optimales Gewicht
Um das DRUGSCORE-Modul mit dem ROSETTA-Modul kombinieren zu können, wurde eine Gewichtsbestimmung mit Testproteinen durchgeführt. Im linken Diagramm wurde die Sequenzidentität (orange) sowie der BLOSUM-Score der Testdesigns gegen die Gewichte aufgetragen, die bei der Suche verwendet worden sind. Die Gewichte sind zur besseren Übersicht mit dem Faktor 50.000 skaliert. Für das gewählte Gewicht von 7.5 / 50000 wurde rechts ein Korrelationsplot erstellt. Hier sind die Sequenzidentitätswerte gegen die BLOSUM-Scores für die einzelnen Proteine des Testdatensatzes aufgetragen.

Wird das Gewicht zu groß gewählt (>30/50000), fällt die Leistung unter das Niveau der Leistung von ROSETTA DESIGN allein. Dann wird der Energiebeitrag durch DRUGSCORE so groß, dass die ROSETTA-Energien an Bedeutung verlieren. Dies bedeutet, dass die Wechselwirkungen der Seitenketten untereinander zugunsten der Wechselwirkungen zum Liganden ignoriert werden. Durch die Wahl eines ungeeigneten Gewichts für DRUGSCORE kann sich also die Leistung von TRANSCENT verschlechtern.

Wird der Testdatensatz unter Verwendung des DRUGSCORE-Moduls und optimiertem Gewicht ausgewertet (Abbildung 34), so ergibt sich folgendes Bild: Für einzelne Testproteine liegt der Sequenzidentitätswert der modellierten aktiven Zentren zwischen 11% und 61% und der Mittelwert liegt bei 37%. Der BLOSUM-Score liegt zwischen -0.6 und 3.7 und im Mittel bei 1.5

Das Einbinden des DRUGSCORE-Moduls mit dem gewählten Gewicht bringt beim Auswerten des Testdatensatzes einen Leistungszuwachs von 5%. Dabei wird die Auswahl derjenigen Reste am stärksten beeinflusst, die Atome in unmittelbarer Nähe zum Liganden haben. Der Einfluss wird mit zunehmendem Abstand schwächer. Da Reste mit einem Abstand bis 7Å frei wählbar sind, unterliegt ein Teil der Reste nur einem schwachen Einfluss, was den moderaten Leistungszuwachs erklärt.

4.3.6 Funktionsdefinition

Für die Evaluierung des dritten Moduls konnte nicht der vollständige Testdatensatz verwendet werden. Testproteine wurden nur verwendet, wenn die Strukturbibliothek mindestens 80 Strukturen umfasst. Diese wurden mit dem Modellierungskriterium „*Sequenzidentität >40% und CORE-Wert >20%*“ generiert (vgl. Kapitel 4.2). Nicht für jedes Testprotein ist die verfügbare Strukturbibliothek ausreichend groß. Für eine zu kleine Strukturbibliothek gibt es drei mögliche Gründe: 1) Es gibt zu dem Protein kein PFAM-Alignment. Das Alignment wird aber benötigt, um aus den Sequenzen Homologiemodelle zu generieren. 2) Das Alignment ist zu klein, hat also weniger als 80 Sequenzen. 3) Das Modellierungskriterium, das die Qualität der Modellierung sicherstellt, filtert zu viele Sequenzen als ungeeignet, so dass zu wenige Sequenzen für Homologiemodelle übrig bleiben. Ausreichend große Strukturbibliotheken konnten nur für 27 Testproteine generiert werden.

Damit die Potentialenergie der Funktionsdefinition und die Energieterme von ROSETTA zusammengeführt werden können, muss die Potentialenergie geeignet gewichtet werden. Dazu wurde der reduzierte Testdatensatz mit verschiedenen Gewichten ausgewertet. Das Intervall, in dem das optimale Gewicht gesucht wurde, liegt zwischen 0 und 16. In diesem Intervall wurden 20 verschiedene Gewichte untersucht. Die Testdesigns wurden 10-mal wiederholt und die Ergebnisse gemittelt (Abbildung 35).

Werden die einzelnen Gewichte gegen die erreichte Sequenzidentität aufgetragen, so beschreiben die Punkte eine Kurve, die bei 30% für ein Gewicht von 0 beginnt und bei 49% bei einem Gewicht von 16 endet. Die Kurve hat ein Maximum von 51% Sequenzidentität bei einem Gewicht von 4. Wenn der BLOSUM-Score auf die gleiche Weise ausgewertet wird, erreicht die Auswertung bei einem Gewicht von 0 einen Wert

von 0.9 und bei einem Gewicht von 16 einen Wert von 2.3. Das Maximum von 2.4 liegt ebenfalls bei einem Gewicht von 4.

Das Maximum der Kurve ist nicht auffällig ausgeprägt. Der Sequenzidentitätswert liegt ab einem Gewicht von 1 bei etwa 50%. Höhere Gewichte haben also kaum negative Auswirkung. Dieser Effekt liegt in der Art der Evaluierung begründet. Ab einem bestimmten Gewicht überwiegt der Einfluss der Funktionsdefinitionsenergie. Reste, die in der Funktionsdefinition beschrieben sind, werden vor allem durch das Funktionsdefinitionsmodul gewählt. Optimal sind vor allem wildtypische Reste, da sie der Funktionsdefinition besonders gut entsprechen. Daher führen auch hohe Gewichte nicht zu einer verschlechterten Leistung.

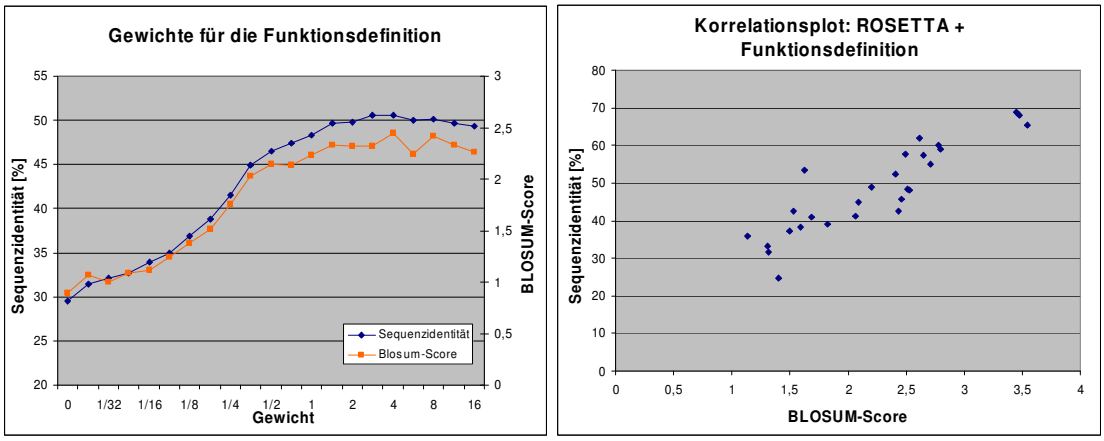


Abbildung 35: Funktionsdefinitions-Modul – Optimales Gewicht
Um das Funktionsdefinitions-Modul mit dem ROSETTA-Modul kombinieren zu können, wurde eine Gewichtssuche mit den 28 Testproteinen durchgeführt. Im linken Diagramm wurde die Sequenzidentität (orange) sowie der BLOSUM-Score der Testdesigns gegen die Gewichte aufgetragen, die bei der Suche verwendet worden sind. Für das gewählte Gewicht von 1 wurde rechts ein Korrelationsplot erstellt. Hier sind die Sequenzidentitätswerte gegen die BLOSUM-Scores für die einzelnen Proteine des reduzierten Testdatensatzes aufgetragen.

Im Anwendungsfall wird auf ein Protein das aktive Zentrum eines anderen Proteins modelliert. In dieser Situation ist die Funktionsdefinition nicht aus dem Protein abgeleitet, auf das sie angewendet wird. Die für die Funktionsdefinition optimalen Reste sind dann selten auch für die gesamte Struktur optimal. Also dürfen die Reste nicht ohne Rücksicht auf die anderen Teilenergien gewählt werden. Um dies zu gewährleisten, sollte das Gewicht für die Funktionsdefinition möglichst klein gewählt sein. Aus diesem Grund wurde ein Gewicht von 1 für die Funktionsdefinition festgelegt (Abbildung 35).

Die Sequenzidentitätswerte für die Testproteine reichten unter diesen Voraussetzungen von 25% bis 69%. Im Mittel liegen sie bei 48%. Der BLOSUM-Score variiert von 1.1 bis 3.5 und liegt im Durchschnitt bei 2.2. Die Sequenzidentität und der BLOSUM-Score korrelieren mit einem Korrelationskoeffizient von 0.90.

Durch den reduzierten Testdatensatz lässt sich der Leistungsgewinn nicht direkt mit den vorhergehenden vergleichen. Da aber die Leistung von ROSETTA DESIGN allein (bei einem Gewicht von 0 für das Funktionsdefinitions-Modul) auf dem reduzierten Testdatensatz (Sequenzidentität: 30%; BLOSUM-Score: 0.9) in etwa der Leistung für den ganzen Testdatensatz (Sequenzidentität: 31%, BLOSUM-Score: 1.1) entspricht, ist der gemessene Leistungszuwachs von ca. 20% plausibel.

4.3.7 PROPKA

Das vierte Modul von TRANSCENT für die Optimierung der pKa-Werte basiert auf der PROPKA-Methode. Um die Energiebeiträge dieses Moduls mit den anderen Energien geeignet kombinieren zu können, muss auch hier ein geeignetes Gewicht gefunden werden. Die Gewichte können aber nicht mit dem ROSETTA-Modul und PROPKA-Modul allein bestimmt werden. Das PROPKA-Modul benötigt das Funktionsdefinitions-Modul, um dynamisch die Reste zu identifizieren, deren pKa-Werte optimiert werden sollen (vgl. Kapitel 4.1.4.1). Damit das Funktionsdefinitions-Modul den Gewichtsfindungsprozess nicht beeinflusst, wird das Gewicht für dieses Modul auf 0 gesetzt. Auf diese Weise wird während der Optimierung die Ähnlichkeit nicht berücksichtigt. Durch die Abhängigkeit vom Funktionsdefinitions-Modul muss der reduzierte Testdatensatz mit nur 27 Proteinen verwendet werden (vgl. Kapitel 4.3.6).

Die pKa-Optimierung erfolgt über einen Strafterm (vgl. Kapitel 4.1.4.2), bei dem sowohl der Vorfaktor *wF* als auch der Exponent *wP* als Parameter wählbar ist. Das Intervall, in dem das optimale Gewicht für *wF* gesucht wurde, liegt zwischen 0 und 8. In diesem Intervall wurden 15 verschiedene Gewichte untersucht. Die Suche wurde für drei verschiedene Gewichte von *wP* wiederholt (*wP*=1.0, *wP*=1.5, *wP*=2.0). Die Testdesigns wurden 10-mal wiederholt und die Ergebnisse gemittelt.

Werden die einzelnen Gewichte gegen die erreichte Sequenzidentität aufgetragen, so beschreiben die Punkte drei Kurven. Alle drei Kurven beginnen bei etwa 29% - 30% bei einem *wF*-Gewicht von 0 und enden bei etwa 38% - 39% für ein *wF*-Gewicht von 8. Keine der Kurven hat ein ausgeprägtes Maximum. Die Maximale Leistung erreicht die Gewichtskombination (*wF*=1.0, *wP*=2.0) mit 39.5% Sequenzidentität.

Wenn der BLOSUM-Score auf die gleiche Weise ausgewertet wird, beginnen die drei Kurven bei etwa 0.9 bei einem *wF*-Gewicht von 0 und enden bei etwa 1.6 für ein *wF*-Gewicht von 8. Auch hier hat keine der Kurven ein ausgeprägtes Maximum. Die Maximale Leistung erreicht die Gewichtskombination (*wF*=1.0, *wP*=2.0) mit 1.7.

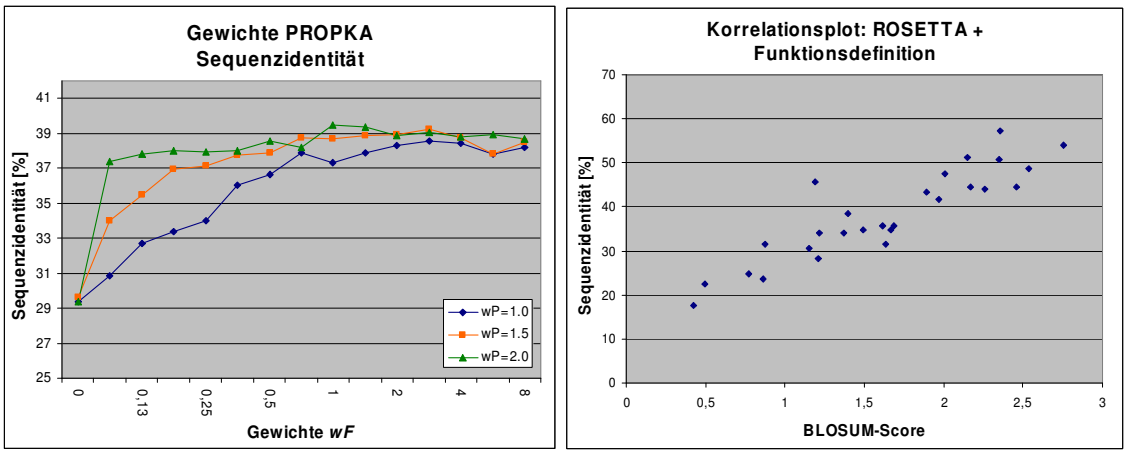


Abbildung 36: PROPKA-Modul – Optimales Gewicht

Um das PROPKA-Modul mit dem ROSETTA-Modul kombinieren zu können, wurde eine Gewichtssuche mit 28 Testproteinen durchgeführt. PROPKA wird als Strafterm der Energiefunktion von TRANSCENT hinzugefügt. Bei der Gewichtssuche wurden verschiedene Vorfaktoren *wF* und Exponentengewichte *wP* für den Strafterm kombiniert. Im linken Diagramm wurde die Sequenzidentität gegen die Gewichte *wF* aufgetragen und die Messung für *wP*=1.0 (blau), *wP*=1.0 (orange) sowie *wP*=2.0 (grün) wiederholt. Für die gewählten Gewichte *wF*=0.5 und *wP*=1.5 wurde rechts ein Korrelationsplot erstellt. Hier sind Sequenzidentitätswerte gegen BLOSUM-Scores für die einzelnen Proteine des Testdatensatzes aufgetragen.

Die Bestimmung eines optimalen Gewichts ist für das PROPKA-Modul aus den gleichen Gründen schwierig, wie beim Funktionsdefinitions-Modul. Auch hier führt eine Überbewertung des Moduls nicht zu einem Leistungsabfall. Die Reste, die das Modul bevorzugt sind eher wildtypisch und vertragen sich energetisch mit den anderen Modulen. Da im Anwendungsfall eine zu hohe Gewichtung des Moduls sehr wohl zu Problemen führen kann, wird auch hier eine Gewichtskombination (*wF*=0.5, *wP*=1.5) gewählt, die bei den Testdesigns zu einer Leistung leicht unterhalb der Maximalleistung führt.

Die Sequenzidentitätswerte für die Testproteine reichen dann von 18% bis 57%. Im Mittel liegen sie bei 38%. Der BLOSUM-Score variiert von 0.4 bis 2.8 und liegt im Durchschnitt bei 1.6. Die Sequenzidentität und der BLOSUM-Score korrelieren mit einem Korrelationskoeffizient von 0.90.

Die deutlich höheren Sequenzidentitätswerte bei Verwendung des PROPKA-Moduls belegen, dass durch die Forderung von optimalen pKa-Werten bestimmte strukturelle Merkmale in aktiven Zentren eindeutig festgelegt sind. Damit bietet die Optimierung von pKa-Werten die Möglichkeit, strukturelle Zusammenhänge in aktiven Zentren zu modellieren, die wahrscheinlich funktionelle Relevanz haben.

4.3.8 Modulkombinationen

Das Grundkonzept für die Modellierungsleistung von TRANSCENT besteht darin, vier Rahmenbedingungen für die Übertragung von aktiven Zentren zu verknüpfen. Diese Rahmenbedingungen werden durch Module repräsentiert: Das ROSETTA-DESIGN- (*R*), das DRUGSCORE- (*D*), das PROPKA- (*P*) und das Funktionsdefinitions-Modul (*F*). Durch die ermittelten Gewichte lassen sich die Module miteinander kombinieren. Dadurch konnte die Leistungsfähigkeit der verschiedenen Modulkombinationen untersucht werden. Von besonderem Interesse war dabei die Kombination, die alle Module beinhaltet.

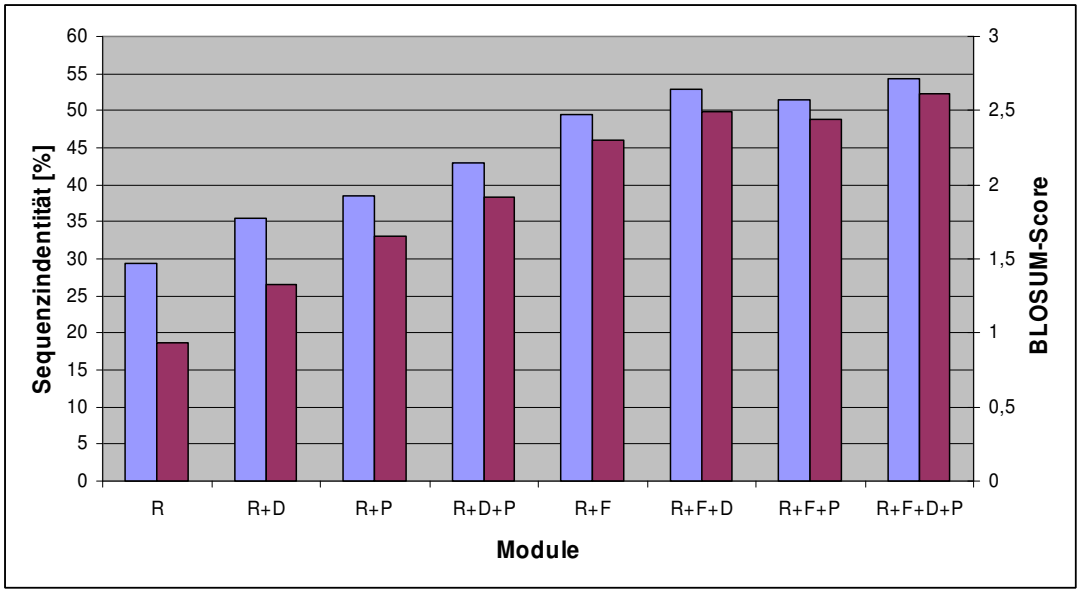


Abbildung 37: Leistungsvergleich der verschiedenen Modulkombinationen
Im Säulendiagramm sind für verschiedene Modulkombinationen die Vorhersageleistungen dargestellt. Für jede Modulkombination wurde der reduzierte Testdatensatz 20-mal ausgewertet und die Ergebnisse gemittelt. Die Säulen in blau zeigen die mittlere Sequenzidentität der Modelle zum Wildtyp. Die Säulen in lila zeigen den gleichen Zusammenhang für den BLOSUM-Score. Die Module sind im einzelnen ROSETTA DESIGN (R), DRUGSCORE (D), Funktionsdefinition (F) und PROPKA (P).

Damit die Leistung der einzelnen Kombinationen miteinander verglichen werden können, wurde in allen Fällen der reduzierte Testdatensatz mit 27 Testproteinen verwendet. Da das *R*-Modul in jedem Fall verwendet werden musste, ergeben sich 8 verschiedene Modulkombinationen. Mit diesen Modulkombinationen wurde der Testdatensatz 20-mal ausgewertet und die Ergebnisse gemittelt (Abbildung 37).

Die mittleren Werte für die Sequenzidentität zwischen Modell und Wildtyp betragen für die verschiedenen Kombinationen: *R* = 20.5%, *R+D* = 35.5%, *R+P* = 38.5%, *R+D+P* = 43.0%, *R+F* = 49.5%, *R+F+D* = 52.8%, *R+F+P* = 51.4% und *R+F+D+P* = 54.3%.

Ähnliche Ergebnisse liefert die Auswertung mit dem BLOSUM-Score: *R* = 0.9, *R+D* = 1.3, *R+P* = 1.6, *R+D+P* = 1.9, *R+F* = 2.3, *R+F+D* = 2.5, *R+F+P* = 2.4 und *R+F+D+P* = 2.6.

Die Ergänzung einer Modul-Kombination um ein zusätzliches Modul führt in allen Fällen im Mittel zu einer Leistungsverbesserung. Allerdings fällt der Leistungsgewinn in den einzelnen Fällen sehr unterschiedlich aus. Während die Kombination *R+F* gegenüber der Kombination *R* eine um 20% höhere Leistung hat, ist der Leistungsgewinn bei *R+P* gegenüber *R* nur 9% und bei *R+D* gegenüber *R* sogar nur 6%. Wenn mehr Module kombiniert werden, fällt der Leistungsgewinn für *D* und *P* noch geringer aus. Er liegt bei *R+F+D* gegenüber *R+F* nur 3.3% höher und bei *R+D* gegenüber *R* sogar nur 1.9%. Dafür ist der Effekt bei der Kombination dieser Module nahezu additiv, *R+F+P+D* gewinnt gegenüber *R+F* 4.8%. Die Gesamtkombination *R+F+P+D* gewinnt gegenüber *R* 24.8%.

Die Verbesserung durch Hinzunahme von Modulen ist in den einzelnen Fällen sehr unterschiedlich. Daher sollte ein einseitiger t-Test klären, ob der Leistungsgewinn immer signifikant ist. Dazu wurden alle Paare von Modulkombinationen (*A*, *B*) untersucht, bei denen *A* aus einer Teilmenge der Module von *B* besteht. Getestet wurde die Nullhypothese „*Die Leistung von Kombination B ist niedriger oder gleich der Leistung von Kombination A*“ gegen die Alternativhypothese „*Die Leistung von Kombination B ist höher als die Leistung von Kombination A*“. Die beiden Stichproben bestanden jeweils aus den mittleren Leistungen der jeweiligen Modulkombinationen für jedes der 27 Proteine des Testdatensatzes (Tabelle 7).

Tabelle 7: t-Test für die Leistungsverbesserung bei Hinzunahme von Modulen

Um die Leistungsverbesserung von Modulkombinationen durch Hinzunahme weiterer Module statistisch zu beurteilen, wurde für jede Variante ein einseitiger t-Test durchgeführt. Die Irrtumswahrscheinlichkeiten wurden in der Tabelle zusammengefasst. Die Nullhypothese lautet jeweils: „*Die Modulkombination <Zeile> ist nicht besser als die Modulkombination <Spalte>*“. Die Alternativhypothese lautet: „*Die Modulkombination <Zeile> ist besser als die Modulkombination <Spalte>*“. Als Stichproben dienten dabei jeweils die Leistungen der Modulkombinationen für die 27 Proteine aus dem Testdatensatz. Die Module sind im einzelnen ROSETTA DESIGN (R), DRUGSCORE (D), Funktionsdefinition (F) und PROPKA (P).

	R	**R+D**	**R+P**	**R+D+P**	**R+F**	**R+F+D**	**R+F+P**
R+D	4,1E-05	-	-	-	-	-	-
R+P	4,5E-08	-	-	-	-	-	-
R+D+P	8,5E-11	1,4E-08	6,3E-05	-	-	-	-
R+F	7,2E-13	-	-	-	-	-	-
R+F+D	4,1E-13	1,8E-12	1,3E-09	1,8E-08	3,4E-04	-	-
R+F+P	1,1E-12	-	1,5E-09	-	0,01	-	-
R+F+D+P	4,0E-13	2,9E-12	9,1E-10	1,0E-09	2,2E-04	0,02	3,6E-03

In allen Fällen, insbesondere auch im Fall von *R+F+D* und *R+F+P* gegenüber *R+F,* kann die Nullhypothese mit einer Irrtumswahrscheinlichkeit von maximal 5%

zurückgewiesen werden. Damit ist die Leistung in allen Fällen signifikant höher, wenn Module zu einer Kombination hinzugenommen werden.

Bei der Analyse von *R* ist die hohe Korrelation der Aminosäureverteilungen zwischen Modell und Wildtyp (Korrelationskoeffizient 90%) aufgefallen (vgl. Kapitel 4.3.4). Auch wenn bei *R* alleine im Mittel nur 30% der Sequenzen im aktiven Zentrum übereinstimmen, waren die relativen Häufigkeiten der Aminosäuren sehr ähnlich. Diese Analyse wurde für die vollständige Modulkombination *R+F+D+P* noch einmal wiederholt. Da für *R+F+D+P* nur der reduzierte Testdatensatz verwendet werden konnte, wurde die Untersuchung auch für *R* mit dem reduzierten Testdatensatz wiederholt (Abbildung 38).

Hierfür wurden insgesamt 727 verschiedene Positionen in den aktiven Zentren der 27 Testproteine untersucht. In beiden Fällen korrelierten die Aminosäureverteilungen von Wildtyp und Modellen im aktiven Zentrum mit einem Korrelationskoeffizient von jeweils 0.86.

Die deutliche Korrelation der Aminosäureverteilungen zeigt, dass sowohl für *R* als auch für *R+F+D+P* die Aminosäurekomposition der Lösungen denen des Wildtyps gleicht.

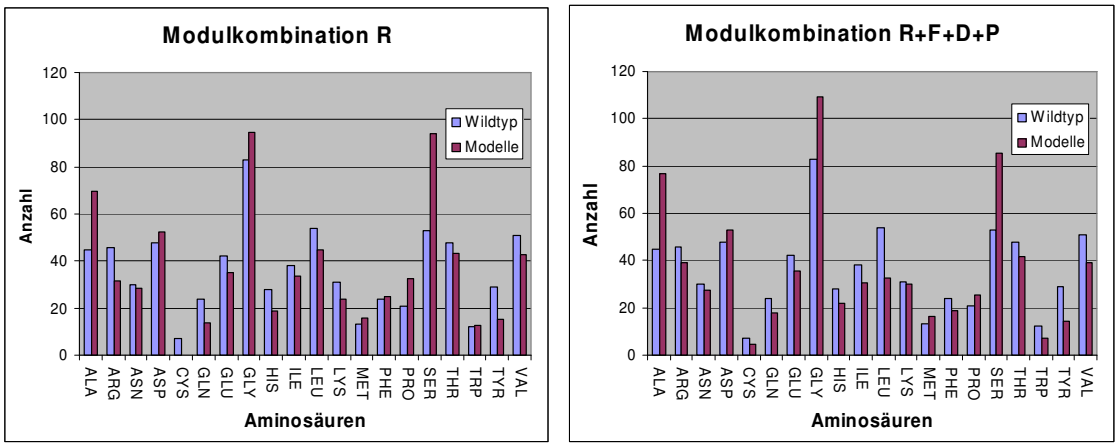


Abbildung 38: Aminosäureverteilungen in den aktiven Zentren von Wildtyp und Modellen
Der reduzierte Testdatensatz besteht aus 27 aktiven Zentren mit insgesamt 727 Aminosäurepositionen. Diese wurden mit dem ROSETTA-Modul (links) und der vollständigen Modulkombination (rechts) neu modelliert. Aufgetragen sind die absoluten Aminosäurehäufigkeiten für die wildtypischen aktiven Zentren (blau) und die Modelle (lila).

Eine weitere Tendenz, die sich bei der Auswertung des Testdatensatzes mit dem *R*-Modul abzeichnete, war die schlechte Performanz für polare und geladene Aminosäuren. Diese wurden im Gegensatz zu den unpolaren Aminosäuren seltener an den richtigen Positionen modelliert. Um diesen Zusammenhang auch für die vollständige Kombination *R+F+D+P* zu untersuchen, wurde diese Auswertung für den reduzierten Testdatensatz mit *R* und *R+F+D+P* wiederholt.

Zur Beurteilung wurde pro Aminosäure ausgewertet, wie häufig sie richtig gewählt wurde (Abbildung 39). Bei *R* wurde am besten Glu (72%) sowie Pro (64%) und am schlechtesten Arg (8%), Asn (6%), Cys (0%), Gln (5%), Glu (10%), His (8%), Met (3%), sowie Tyr (8%) vorhergesagt. Die Standardabweichung für die Sequenzidentität pro Aminosäure beträgt 20% bei einem Mittelwert von 29%.

Bei der Kombination *R+F+D+P* wurden am besten Gly (86%), Pro (72%) und Ser (70%) und am schlechtesten Tyr (27%) und Met (25%) vorhergesagt. Die Standardabweichung für die Sequenzidentität pro Aminosäure beträgt 15% bei einem Mittelwert von 54%.

Ebenso wie für den vollständigen Testdatensatz hat das *R*-Modul alleine vor allem Schwierigkeiten bei der Vorhersage von polaren und geladenen Aminosäuren. Diese Tendenz ist bei der Gesamtkombination nicht mehr zu beobachten. Die Sequenzidentität pro Aminosäure entspricht für *R+F+D+P* deutlicher der Gesamtleistung als bei *R*.

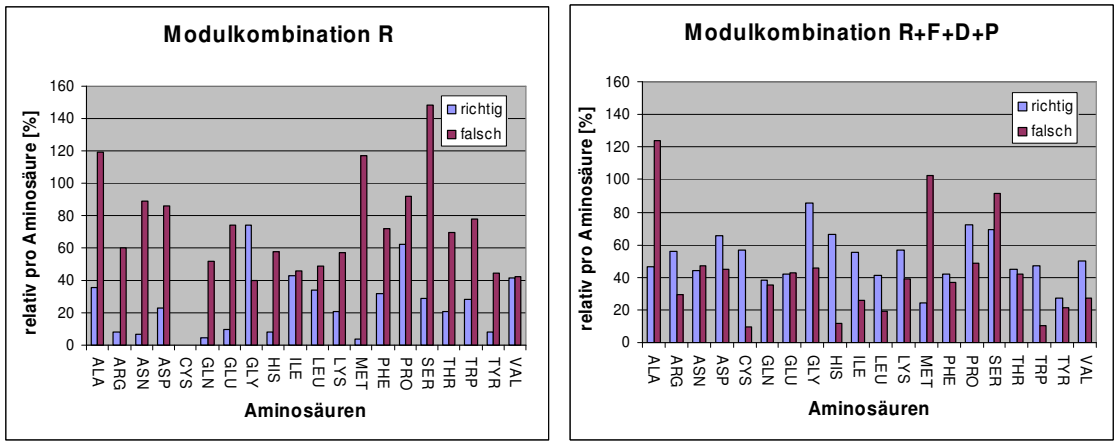


Abbildung 39: Sequenzidentität pro Aminosäure

Für die aktiven Zentren des reduzierten Testdatensatzes wurde pro Aminosäure ausgewertet, wie häufig sie richtig gewählt wurden. Die Auswertung wurde mit dem ROSETTA-Modul allein (links) und der vollständigen Modulkombination (rechts) durchgeführt und 20-mal wiederholt. In den Diagrammen ist zum einen pro Aminosäure die gemittelte Häufigkeit von richtigen ausgewählten Fällen aufgetragen (blau). Zum anderen ist aufgetragen, wie oft - relativ zum wildtypischen Vorkommen - die Aminosäure an einer falschen Position gewählt worden ist (lila). Mit *R* allein wurden 30% der wildtypischen Serine richtig gesetzt. Gleichzeitig wurden 150% mehr Serine falsch ausgewählt als in den wildtypischen aktiven Zentren vorkamen.

Jeder Aminosäuretyp profitierte von der Leistungssteigerung, da die Aminosäure häufiger richtig vorhergesagt wurde. Die meisten Aminosäuren profitierten zusätzlich davon, dass die Aminosäure seltener an einer falschen Stelle vorhergesagt wurde (Abbildung 39). Auf diese Weise steigt die bedingte Wahrscheinlichkeit, dass eine Aminosäure dem Wildtyp entspricht, wenn sie an einer konkreten Stelle im Modell vorliegt (Abbildung 40). Diese Wahrscheinlichkeit wird aus dem Verhältnis zwischen der Häufigkeit, dass eine Aminosäure richtig gewählt wurde und der Häufigkeit, dass eine

Aminosäure insgesamt gewählt wurde berechnet und lässt sich als Modellierungssicherheit interpretieren.

Bei der alleinigen Verwendung des Moduls *R* ist die Auswahl von Gly (65%), Val (49%) und Ile (48%) am sichersten. Am unsichersten ist die Auswahl von Met (3%), Asn (7%) und Gln (8%). Da Cys nie gewählt wurde, kann es in diesem Zusammenhang nicht beurteilt werden.

Für die vollständige Modulkombination sind die am sichersten ausgewählten Aminosäuren Cys (86%), His (85%) und Trp (82%). Am unsichersten sind die Aminosäuren Met (19%) und Ala (23%).

Wären die wildtypischen Aminosäuren gleichverteilt, so würde bei einer zufallsbasierten Modellierung jede Aminosäure mit 5% Wahrscheinlichkeit richtig gewählt. Bei der Gesamtkombination liegt die Wahrscheinlichkeit bei jeder Aminosäure über diesem Wert und auch über dem Wert des R-Moduls allein.

Obwohl die Vorhersagen von *R+F+D+P* im Mittel in 54% der Fälle richtig liegen, schwankt die Leistung für Vorhersage der einzelnen Aminosäuren. Diese Verzerrung sollte im Anwendungsfall berücksichtigt werden.

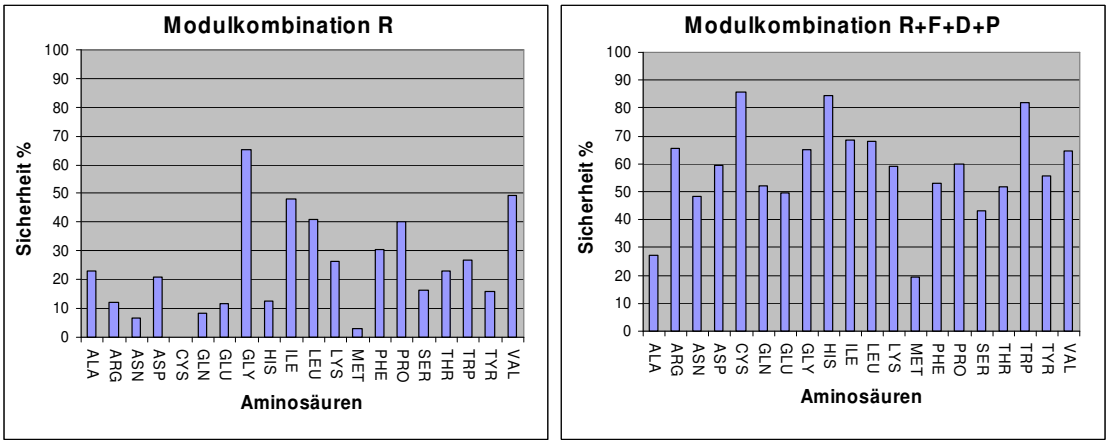


Abbildung 40: Bedingte Wahrscheinlichkeit für die richtige Aminosäurewahl
Die Wahrscheinlichkeit, dass eine gewählte Aminosäure richtig ist, lässt sich für jeden Aminosäurentyp als Verhältnis zwischen den Häufigkeiten „richtig gewählt“ und „insgesamt gewählt“ ausdrücken. Links sind diese Wahrscheinlichkeiten für das ROSETTA Modul allein angegeben und rechts für die vollständige Kombination aller Module. Da Cystein durch das ROSETTA-Modul allein nie gewählt wurde, wird es in der linken Auswertung nicht berücksichtigt.

Es ist nicht einfach, die Leistung der Module in Bezug auf die einzelnen Aminosäuren zu beurteilen. So werden zum Beispiel wildtypische Cysteine in 57% der Fälle gefunden, was der Gesamtleistung der Kombination ungefähr entspricht. Cysteine werden aber nur selten an falschen Stellen modelliert, womit ein Cystein im Modell zu 86% auch im

Wildtyp zu finden ist. Da Cysteine insgesamt aber selten vorkommt (7 Fälle ~ 1%), fällt es bei der Gesamtleistung kaum ins Gewicht.

Die Gesamtkombination sagt zwar nicht die Belegung aller Positionen richtig voraus, aber eine derartige Leistung ist für den Transfer von aktiven Zentren auch nicht unbedingt erforderlich. Da aktive Zentren nicht vollständig konserviert sind, kann davon ausgegangen werden, dass nicht alle Aminosäuren in gleichem Maß relevant sind. Zusätzlich ist das Abstandskriterium bei der Definition des aktiven Zentrums mit 7Å zum Liganden sehr großzügig formuliert. Nicht alle Reste in diesem Abstand haben eine Funktion für das aktive Zentrum. Um zu überprüfen, ob die Vorhersageleistung vom Konservierungsgrad der Aminosäuren abhängig ist, wurde für jede Modulkombination die mittlere Konserviertheit der richtig und der falsch vorhergesagten Positionen ermittelt (Abbildung 41).

Die durchschnittliche Konserviertheit liegt bei den betrachteten Resten im reduzierten Testdatensatz bei 76%. Wird für die einzelne Modulkombinationen die Konserviertheit der richtig vorhergesagten Reste beurteilt, beträgt sie für *R* alleine 79% und für *R+F+D+P* 84%. Der Konserviertheitsgrad wird aber nicht in jedem Fall durch Hinzunahme weiterer Module gesteigert. Während *R+D* (81%), *R+P* (82%) und *R+F* (83%) gegenüber *R* jeweils höhere Konserviertheit aufweisen, steigert sich für *R+F+D* (83%) gegenüber *R+F* (83%) und *R+F+D+P* (84%) gegenüber *R+F+P* (84%) die Konserviertheit nicht. Bei *R+F+P* (84%) gegenüber *R+F* (83%) und auch bei *R+F+D+P* (84%) gegenüber *R+F+D* (83%) steigert sich die Konserviertheit jedoch jeweils um etwa 1%. Die durchschnittliche Konserviertheit der falsch vorhergesagten Aminosäurepositionen sinkt für die einzelnen Modulkombinationen im gleichen Verhältnis von 75% für R auf 67% für *R+F+D+P*.

Wird für die Modulkombinationen auf ähnliche Weise der mittlere Abstand ausgewertet (Abbildung 41), ergibt sich ein mittlerer C_β-Abstand (für Gly C_α-Abstand) zum nächsten Ligandenatom von durchschnittlich 3.9Å. Dieser Abstand sinkt für die Menge der richtig vorhergesagten Positionen sukzessiv von 4.0Å für *R* auf 3.6Å für *R+F+D+P*. Für die falsch vorhergesagten Positionen steigt der Abstand von 3.9Å für *R* auf 4.3Å für *R+F+D+P*.

Für das R-Modul unterscheiden sich die Konserviertheit und der mittlere Abstand für richtig und falsch ausgewählte Reste kaum. Sie liegen jeweils in der Nähe des Mittelwerts. Dagegen liegen die Werte bei der Gesamtkombination deutlich auseinander. Das lässt den Schluss zu, dass Reste mit einem Bezug zur Funktion häufiger richtig ausgewählt werden.

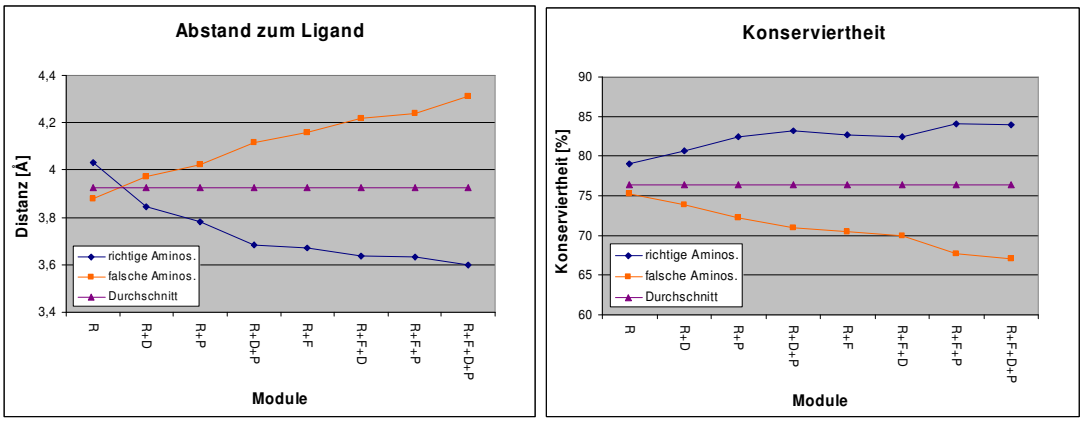


Abbildung 41: Ligandenabstand und Konserviertheit der wieder gefundenen Reste
Das linke Diagramm beschreibt den Einfluss der verschiedenen Modulkombinationen auf den mittleren Abstand zum Liganden. Die blaue Kurve beschreibt die Mittelwerte der richtig modellieren Reste, die orange Kurve die Mittelwerte der falsch modellierten Reste und die lila Kurve zeigt die Mittelwerte für alle Reste der wildtypischen aktiven Zentren. Im rechten Diagramm wird der Einfluss der verschiedenen Module auf die Konserviertheit beschrieben. Die blaue Kurve zeigt wieder die Mittelwerte für die korrekt modellierten Reste, die orange die Mittelwerte für die falschen und die lila Kurve zeigt die Mittelwerte für alle Reste der wildtypischen aktiven Zentren. Die Module sind im einzelnen ROSETTA DESIGN (R), DRUGSCORE (D), Funktionsdefinition (F) und PROPKA (P).

Als letzte Fragestellung wurde die Abhängigkeit der Leistung von der Gesamtenergie untersucht. In TRANSCENT wird SA als Optimierungsverfahren verwendet. Da beim SA durch Wiederholung verschiedene Lösungen gefunden werden können, bietet es sich an, zu untersuchen, ob die Güte der Lösung mit der Energie korreliert. Für jede Modulkombination wurde daher die folgende Auswertung durchgeführt: Für jedes Protein aus dem reduzierten Testdatensatz wurden 20 Modelle generiert. Anschließend wurden die Sequenzidentitätswerte für die Lösung mit der höchsten und der niedrigsten Energie sowie die durchschnittliche Übereinstimmung berechnet. Diese Ergebnisse wurden über die 27 Testproteine gemittelt.

In allen Fällen war die Sequenzidentität bei höherer Energie geringer als bei niedrigerer Energie. Die Unterschiede bewegen sich zwischen 0.9% für R+F+D und 4.4% für R+F+D+P. Die mittlere Sequenzidentität liegt erwartungsgemäß jeweils zwischen den Werten für hohe und niedrige Energie.

Der Unterschied zwischen den Lösungen mit den jeweils schlechtesten und besten Energiewerten fällt vor allem bei der *R+F+D+P* Kombination sehr deutlich aus. Daher ist es sinnvoll, durch mehrfaches Wiederholen die Lösungsmenge zu vergrößern. Da im Anwendungsfall diese Lösungen nicht einfach gemittelt werden können, ist es empfehlenswert die Variante mit niedrigster Energie auszuwählen.

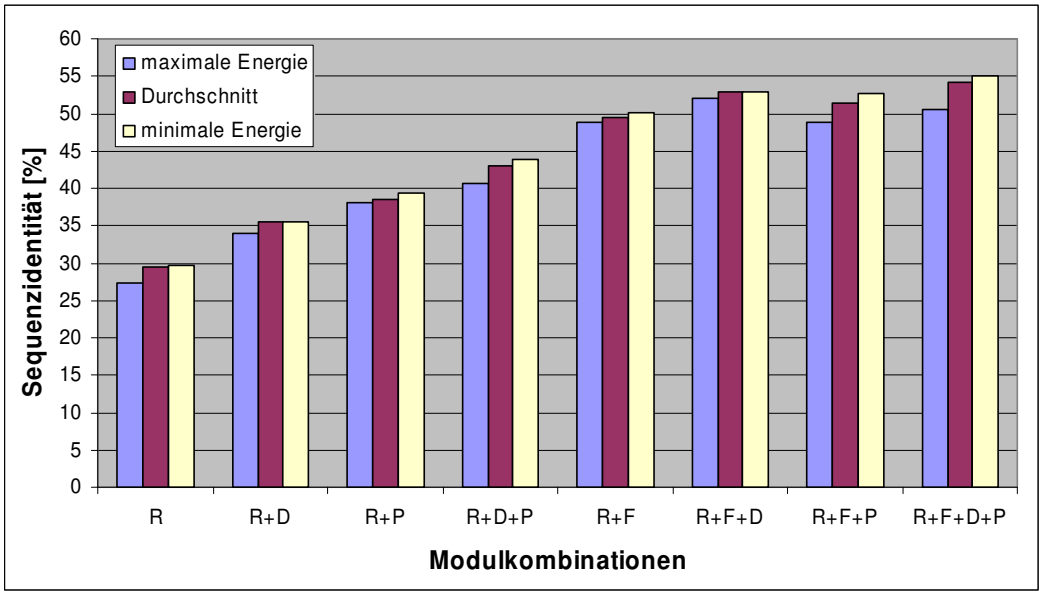


Abbildung 42: Leistungsunterschiede in Abhängigkeit von der Energie
In TRANSCENT wird SA als Optimierungsverfahren verwendet. Um zu untersuchen, ob sich die Qualität der gefundenen Lösungen durch die Energie beurteilen lässt, wurden für verschiedene Modulkombinationen der reduzierte Testdatensatz 20-mal ausgewertet und die Sequenzidentität ermittelt. Dabei wurden einmal die Lösungen mit höchster Energie gemittelt (blau) und einmal die Lösungen mit niedrigster Energie (gelb). Zusätzlich wurden alle Lösungen gemittelt (rot). Die Module sind im einzelnen ROSETTA DESIGN (R), DRUGSCORE (D), Funktionsdefinition (F) und PROPKA (P).

4.3.9 Beispiel Oxidoreduktase Cytochrom P450

Um zu veranschaulichen, welchen Einfluss die Module im Detail auf den Optimierungsprozess haben, werden nun die Ergebnisse an einem konkreten Beispiel diskutiert: Der Struktur der Oxidoreduktase Cytochrom P450 2B4 aus *Oryctolagus cuniculus* (Kaninchen). Im aktiven Zentrum ist eine Häm-Gruppe gebunden, die als Kofaktor den Oxidoreduktionsprozess ermöglicht (Abbildung 43). Pro Modulkombination wurde das Testdesign mit der Oxidoreduktase 20-mal wiederholt und die jeweils energetisch beste Lösung gewählt.

Im aktiven Zentrum wurden 39 Reste als wählbar definiert, von denen das *R*-Modul allein 26% wieder finden konnte. Die Gesamtkombination erreichte 59%. Auch in diesem Beispiel steigt die Leistung durch Hinzunahme von Modulen. Die Präferenzen der einzelnen Module lassen sich in einem MSA gut analysieren, in dem nur die Positionen des aktiven Zentrums aufgeführt sind (Abbildung 44). So wird eine Gruppierung von drei Arg-Resten (Pos. 2,5 und 30 im MSA in Abbildung 44) durch *R* und *R+D* nicht gefunden. Durch *R+F* und *R+F+D* werden zwei gefunden und anstelle des dritten Arginins mit einem Glu zu einer Salzbrücke kombiniert. Erst durch Modulkombinationen mit *P* werden alle drei gefunden.

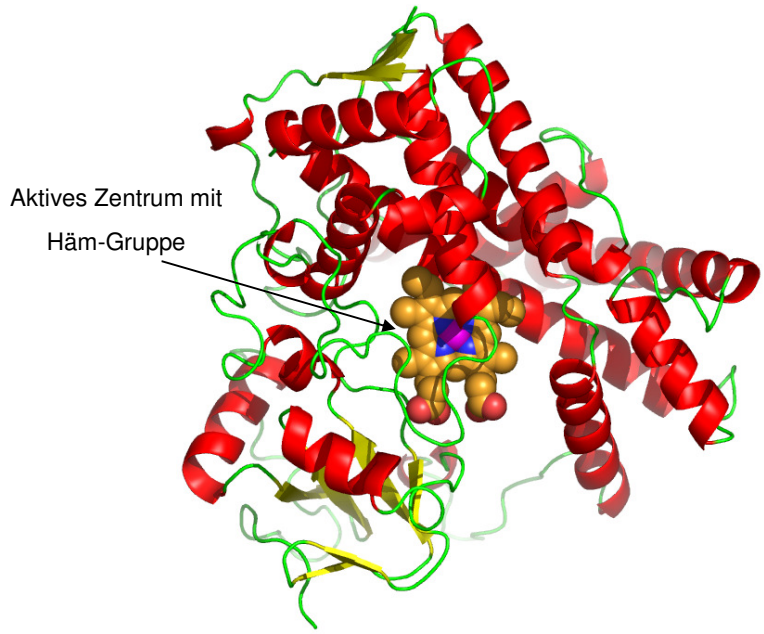


Abbildung 43: Struktur der Oxidoreduktase Cytochrom P450 2B4
Zu sehen ist die Struktur 1po5 von Oxidoreduktase Cytochrom P450 2B4 aus dem Kaninchen in Bänderdarstellung. Der Ligand im aktiven Zentrum ist eine Häm-Gruppe (Kugeldarstellung), dessen zentrales Eisen (rosa) Elektronen aufnehmen und abgeben kann.

An einer anderen Stelle (Pos. 20 im MSA) wird ein wildtypisches Val von *R, R+F* sowie *R+F+P* gefunden, alle anderen Kombinationen und damit auch *R+F+D+P*, wählen aber ein Tyr oder Phe. Hier führt der Kompromiss zu einer falschen Auswahl.

	10	20	30	Richtig
R	LTWLKVDMAKGSDAALDNGVPFALNPFGSLADLFAMRML			10 (26%)
R+D	DQEEAVKLAWGSPTALIFGYPHSLNPFGSHASRGAQAML			12 (31%)
R+P	DRHERADIARASDVALDNGYPHSLNPFGSRACEFAIAEL			14 (36%)
R+D+P	DRELRVKMAWGSPTALIFGFPHSLNPFGSRACEGAIAEL			17 (44%)
R+F	DRELEVDMALGSTTTQALGVPHSLNPDSSRSCPFAMIML			16 (41%)
R+F+D	DRELEVLMAWGSTTTQQMGYPHSLNPFSSRACLGATAML			19 (49%)
R+F+P	NRLLRVDMALGSTTTQALGVPHSLNPTSSRQCLFAIIEL			21 (54%)
R+F+D+P	NRLLRVLMAWGSTTTQQMGFPHSLNPFSSRACLGAIAEL			23 (59%)
1po5	LRLMRISLFAGTTTTQLIGVPHPLMPFSLRICLGEIAEL			39 (100%)

Abbildung 44: MSA der verschiedenen Modelle und des Wildtyps von Cytochrom P450
Das Testprotein Oxidoreduktase Cytochrom P450 2B4 wurde mit allen Modulkombinationen 20-mal ausgewertet. Für jede Kombination wurde die Lösung mit der niedrigsten Energie ausgegeben. Die Sequenzen der Lösungen wurden in einem MSA zusammengefasst. In der Abbildung sind nur die Positionen des MSAs zu sehen, die bei der Optimierung frei wählbar waren. Die unterste Sequenz ist die des Wildtyps. Links von jeder Sequenz ist die Modulkombination aufgeführt und rechts die Übereinstimmung (absolute Anzahl und relativer Anteil bezogen auf die Größe des aktiven Zentrums) mit der Wildtyp-Sequenz. Identische Aminosäuren sind blau und ähnliche hellblau hinterlegt. Die Module sind im einzelnen ROSETTA DESIGN (*R*), DRUGSCORE (*D*), Funktionsdefinition (*F*) und PROPKA (*P*).

In den meisten Fällen lässt sich aber für *R+F+D+P* eine Tendenz zum wildtypischen Konsensus beobachten. Nur in 2 von 23 Fällen (Pos. 1 und 20) wurde eine wildtypische Auswahl, die in einer Teilkombination gefunden wurde, in R+F+D+P wieder verworfen.

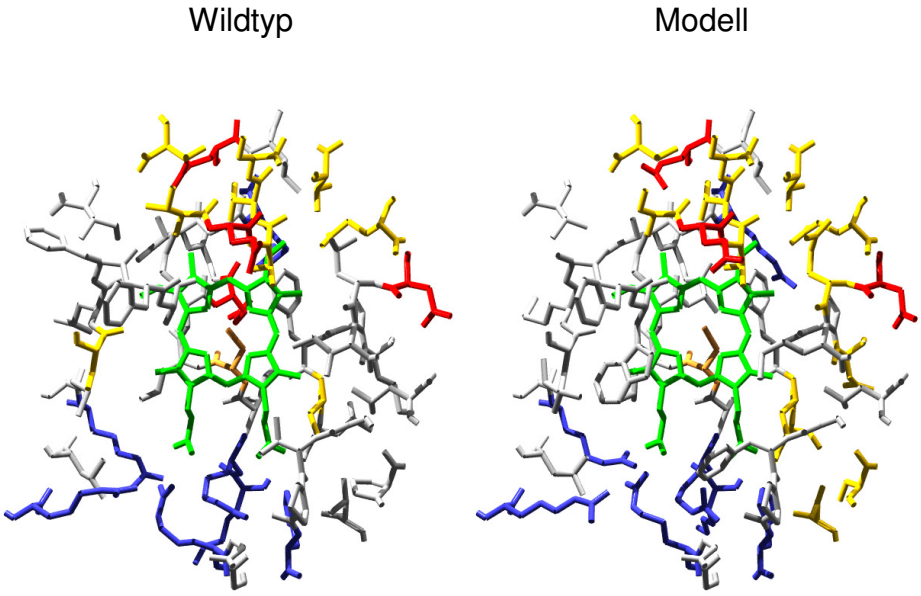


Abbildung 45: Vergleich der aktiven Zentren von Wildtyp und bestem Modell
Links ist die Struktur des aktiven Zentrums des Testproteins Oxidoreduktase Cytochrom P450 2B4 mit gebundener Häm-Gruppe (grün) zu sehen. Recht ist das aktive Zentrum des energetisch besten Modells zu sehen. Das Modell wurde mit der Kombination R+F+D+P generiert. Die Färbung zeigt Arg, His sowie Lys in blau. Asp und Glu sind rot eingefärbt. Asn, Cys, Gln, Ser, Thr sowie Tyr sind gelb gefärbt. Die hydrophoben Reste sind grau dargestellt.

Werden die Strukturen des wildtypischen aktiven Zentrums mit den Modellen verglichen, so zeigt sich die Leistung der *R+F+D+P*-Kombination auch auf struktureller Ebene (Abbildung 45). Der RMSD für die Seitenkettenatome der 23 richtig gewählten Reste beträgt nur 1.0Å.

Mit diesem Beispiel soll nicht behauptet werden, dass die Methode besonders geeignet sei aktive Zentren von Oxidoreduktasen zu transferieren. Dies ist schon deswegen schwierig, weil neben der Häm-Gruppe (Kofaktor) das eigentliche Substrat in der Struktur nicht vorhanden ist (Scott et al., 2003). Trotzdem zeigt das Beispiel, bis zu welchem Grad die Methode in der Lage ist, ein aktives Zentrum zu beschreiben und zu modellieren. Es zeigt außerdem, wie sich die einzelnen Module dabei gegenseitig ergänzen.

4.4 Evaluierung mit Ribulosephosphat-bindenden $(\beta\alpha)_8$-Barreln

Für eine weitere Evaluierung von TRANSCENT, wurden Umwandlungsdesigns zwischen fünf Enzymen der Ribulosephosphat-bindenden $(\beta\alpha)_8$-Barrel-Superfamilie modelliert: HisA, HisF, TrpA, TrpC sowie TrpF. Gleichzeitig sollte mit den Umwandlungen die Arbeitsweise des Programms im Anwendungsfall erläutert werden. Zusätzlich stellen die Umwandlungsmodellierungen konkrete Vorhersagen dar, die beschreiben mit welchen Austauschen sich ein neues aktives Zentrum etablieren lässt.

4.4.1 Umwandlungskombinationen

Um die Umwandlungen durchführen zu können, werden für jedes Protein eine Struktur mit Liganden benötigt (vgl. Kapitel 3.3.2). Die hier verwendeten Strukturen sind in Tabelle 8 aufgeführt.

Der Ligand für HisA und HisF ist jeweils PRFAR, das Produkt von HisA und Substrat von HisF. Der Ligand von TrpA ist IGP, das ist das Substrat der TrpA-Reaktion. In TrpF und TrpC ist jeweils rCdRP gebunden. Der rCdRP-Ligand ist sowohl das Produkt-Analogon der TrpF-Reaktion als auch das Substrat-Analogon der TrpC-Reaktion.

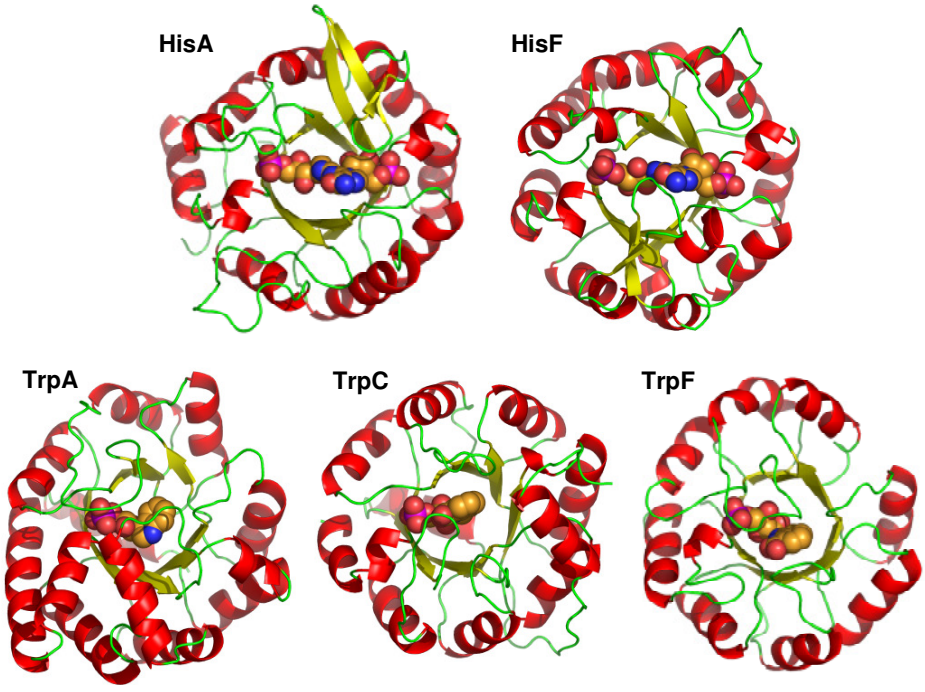


Abbildung 46: Strukturen von fünf Ribulosephosphat-bindenden $(\beta\alpha)_8$-Barrel-Enzymen
Die Abbildung zeigt die Strukturen in Bänderdarstellung mit gebundenen Liganden in Kugeldarstellung. Die Strukturen sind im Einzelnen: HisA mit PRFAR, HisF mit PRFAR, TrpA mit IGP, TrpC mit rCdRP sowie TrpF mit rCdRP. Alle Liganden weisen einen Phosphat-Rest auf, der an äquivalenten Positionen der jeweiligen Strukturen gebunden ist. Die Liganden von HisA und HisF sind nahezu doppelt so groß wie die Liganden von TrpA, TrpF und TrpC.

Tabelle 8: Beschreibung der Strukturen von HisA, HisF, TrpA, TrpC und TrpF
In der Tabelle sind Strukturen der Proteine aufgelistet, die für die Umwandlungsexperimente verwendet worden sind. Für jedes Protein sind der Name, der Organismus aus dem das Protein stammt, der PDB-Code der Kristallstruktur und der gebundene Ligand angegeben.

Proteinname	Organismus	PDB-Code	Ligand
HisA	*T. maritima*	1qo2	PRFAR (modelliert)
HisF	*T. maritima*	1thf	PRFAR (modelliert)
TrpA	*S. typhimorium*	1qoq	IGP
TrpC	*S. solfataricus*	1lbf	rCdRP
TrpF	*T. maritima*	1lbm	rCdRP

Obwohl alle Strukturen die typische strukturelle Charakteristik der $(\beta\alpha)_8$-Barrel-Faltung aufweisen (Abbildung 46), sind lokale Unterschiede vor allem in den Schleifenregionen der aktiven Zentren vorhanden. Der MULTIPROT-Server (Shatsky et al., 2004) gibt als C_α–RMSD zwischen den fünf Strukturen einen Wert von 1.7Å an. Bei dieser Berechnung werden aber nur 75 Positionen berücksichtigt, die in allen Strukturen ähnlich sind. Wenn dagegen alle Positionen berücksichtigt werden, die einander zugeordnet werden können, so liegt der Wert bei 3.0Å. In Tabelle 9 sind C_α–RMSD-Werte über alle Positionen für Strukturpaare angegeben, diese wurden mit TM-ALIGN (Zhang & Skolnick, 2005) berechnet.

Tabelle 9: Strukturelle Unterschiede zwischen HisA, HisF, TrpA, TrpC und TrpF
Es sind die C_α–RMSD-Werte zwischen allen Paarkombinationen aufgeführt. Berücksichtigt wurden dabei alle C_α-Abstände von Positionspaaren, die über ein strukturbasiertes Sequenzalignment einander zugeordnet worden sind.

	HisA	HisF	TrpA	TrpC	TrpF
HisA	-	3.1Å	3.5Å	3.0Å	2.8Å
HisF	3.1Å	-	2.6Å	3.3Å	2.8Å
TrpA	3.5Å	2.6Å	-	3.2Å	3.1Å
TrpC	3.0Å	3.3Å	3.2Å	-	2.5Å
TrpF	2.8Å	2.8Å	3.1Å	2.5Å	-

Bei den Testdesigns (vgl. Kapitel 4.3.1) sind die Proteingerüste der Modelle identisch zu den Proteingerüsten der Vorlage. Bei den Umwandlungsdesigns unterscheidet sich der Rückgratverlauf der Proteingerüste von denen der Vorlage. Um diesen Effekt zu kompensieren, lässt sich in TRANSCENT eine Mindestvarianz für die Wahrscheinlichkeitsverteilungen angeben. Diese Mindestvarianz wurde für die Umwandlungen auf 2Å festgelegt.

Werden alle Kombinationen von Umwandlungen berücksichtigt, so ergeben sich 25 verschiedene Experimente. Fünf davon sind trivial, weil dann Vorlage und Gerüst identisch sind. Diese Umwandlungen entsprechen dem Konzept des Testdesigns und können Aufschluss darüber geben, wann und in welchem Umfang ein übertragenes aktives Zentrum im Idealfall gleichartig ist. Daher wurde mit TRANSCENT für die fünf Enzyme 10-mal ein Testdesign durchgeführt und gemittelt (Tabelle 10).

Tabelle 10: Sequenzidentitätswerte für einzelne Umwandlungsexperimente
Zwischen HisA, HisF, TrpA, TrpC sowie TrpF wurden alle Kombinationen von Umwandlungen zwischen HisA, HisF, TrpA, TrpC sowie TrpF berechnet. Für diese Tabelle wurden die Sequenzidentitätswerte der Umwandlungen zum wildtypischen Proteingerüst bestimmt. Hervorgehoben sind die Sequenzidentitätswerte von Testdesigns, bei denen das aktive Zentrum auf das Proteingerüst modelliert wurde, das auch als Vorlage diente.

	HisA	HisF	TrpA	TrpC	TrpF
HisA-Aktivität	**46%**	37%	35%	34%	29%
HisF-Aktivität	39%	**48%**	21%	23%	28%
TrpA-Aktivität	15%	27%	**54%**	26%	18%
TrpC-Aktivität	20%	20%	22%	**53%**	17%
TrpF-Aktivität	30%	32%	30%	33%	**53%**

Die Sequenzidentitätswerte für die Testdesignexperimente sind im Einzelnen: 46% für HisA, 48% für HisF, 54% für TrpA, 53% für TrpC sowie 53% für TrpF. Damit liegt die Leistung für die Enzyme HisA und HisF leicht unterhalb der mittleren Leistung des Programms (54% für den Testdatensatz). Die Leistung für TrpA, TrpC und TrpF entspricht ziemlich genau der mittleren Leistung des Programms. Also kann angenommen werden, dass die Methode in der Lage ist, die funktionalen Zusammenhänge in den jeweiligen aktiven Zentren im erwarteten Grad zu erfassen

Wird diese Auswertung für die anderen 20 Kombinationen durchgeführt, liegt der mittlere Sequenzidentitätswert für die einzelnen Fälle zwischen 17% für TrpA-Aktivität auf HisA und 39% für HisF-Aktivität auf HisA. Im Mittel liegt die Sequenzidentität bei 27% (Tabelle 10).

Erwartungsgemäß ist in diesen Fällen die Sequenzidentität niedriger, da auf den Proteingerüsten nicht die wildtypischen Aktivitäten etabliert wurden. Identisch gewählte Aminosäuren werden wahrscheinlich vor allem durch Beschränkungen des Proteingerüsts festgelegt. Dafür spricht auch der Mittelwert von 27%, der in etwa der durchschnittlichen Leistung entspricht, wenn das ROSETTA-Modul allein verwendet wird. Die Werte geben zwar keinen Aufschluss darüber, wie gut die Umwandlung modelliert ist, aber spiegeln den Aufwand wider, der notwendig sind, um die neue Funktion zu etablieren. In Fällen

mit besonders niedriger Sequenzidentität wird die neue Funktion durch entsprechend viele Austausche modelliert.

4.4.2 Analyse der Energiebeiträge

Auch wenn sich die Sequenzidentitätswerte in den erwarteten Bereich bewegen, ist es kaum möglich, aus den Werten den Erfolg der Umwandlungsmodellierungen abzuschätzen. Ein besser geeigneter Indikator ist die Analyse der Energiebeiträge, die durch die vier Module von TRANSCENT für die jeweils betrachtete Lösung berechnet wurden. Diese Energiebeiträge spiegeln wider, in welchem Grad die einzelnen Rahmenbedingungen erfüllt sind (siehe Abbildung 47 und Abbildung 48). Bei der Interpretation müssen allerdings die Charakteristika der einzelnen Module berücksichtigt werden. Generell können die Ergebnisse der jeweiligen Testdesigns als Referenz herangezogen werden.

Das ROSETTA-Modul bewertet die Stabilität eines Proteins ganzheitlich, d. h. nicht auf das aktive Zentrum allein bezogen. Um den Einfluss des neuen aktiven Zentrums auf die Stabilität zu erfassen, muss die Stabilität eines umgewandelten Proteins relativ zur Stabilität des Proteingerüsts betrachtet werden. Als Energiebeitrag für die Stabilität des Grundgerüsts wird die Energie des entsprechenden Testdesigns angenommen.

Das DRUGSCORE-Modul bewertet die Bindungsaffinität zwischen dem Proteinmodell und dem Liganden. Negative Energiebeiträge implizieren dabei gute Bindung und positive Energiebeiträge entsprechend schlechte Bindung. Maßgeblich für die Beurteilung der Eignung der Bindung ist aber nicht nur das Vorzeichen des Energiebeitrags, sondern auch der Unterschied des Energiebeitrags im Vergleich zum Testdesign. Durch das Testdesign ist der Bindungsenergiebeitrag bei optimalen Voraussetzungen gegeben.

Die Interpretation der Energiebeiträge des Funktionsdefinitionsmoduls ist analog zur Interpretation des DRUGSCORE-Moduls. Niedrige Beiträge deuten auf eine ähnliche HB-Gruppenverteilung wie in der Vorlage hin. Auch hierbei sollte das Vorzeichen des Energiebeitrags und der Unterschied zum Energiebeitrag für das entsprechende Testdesign berücksichtigt werden.

Da der Energiebeitrag des PROPKA-Moduls als Strafterm realisiert ist, kann er nur positive Werte annehmen. Optimale pKa-Werte sind bei entsprechend kleinen Energiebeiträgen des Moduls erreicht. Auch hier liefert der Energiebeitrag der entsprechenden Testdesigns Hinweise auf Probleme bei der Optimierung der pKa-Werte.

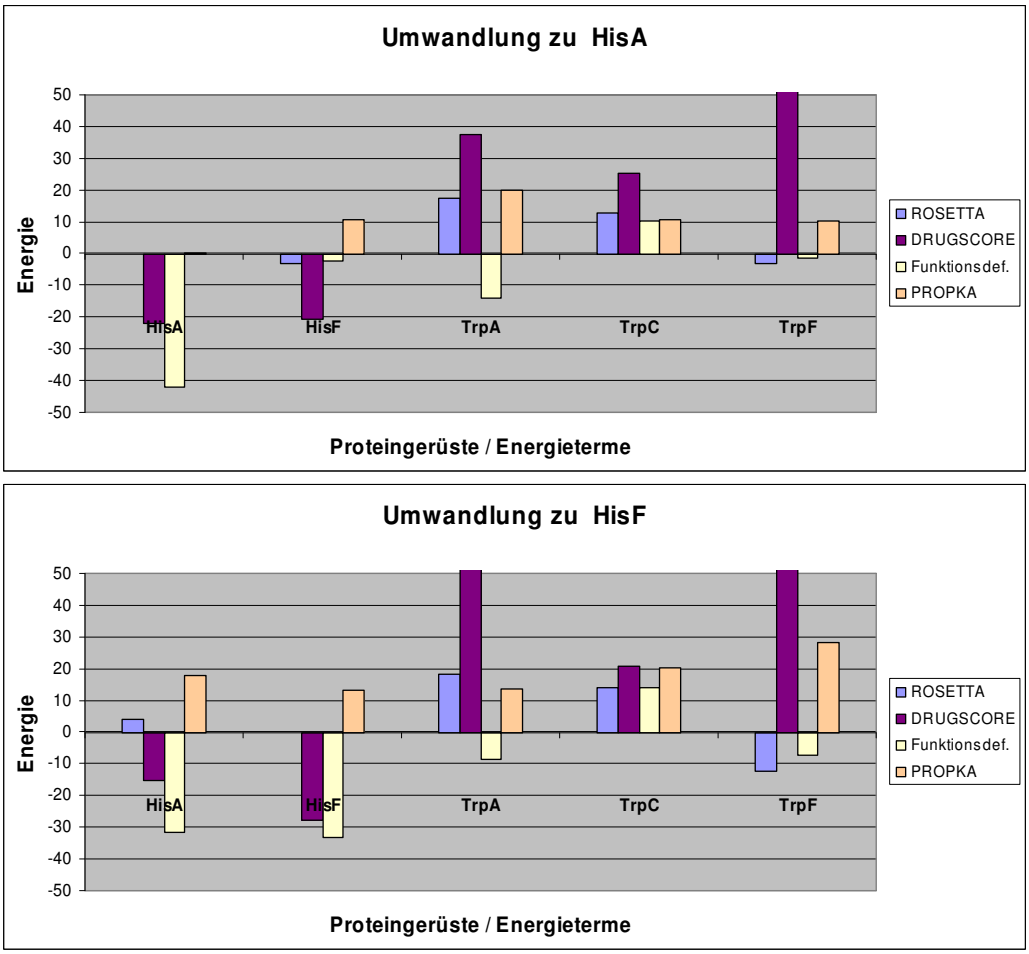


Abbildung 47: Energieprofile für HisA und HisF Umwandlungen

Die aktiven Zentren von HisA (oberes Diagramm) sowie HisF (unteres Diagramm) wurden jeweils auf fünf Proteingerüste übertragen. Dies sind HisA, HisF, TrpA, TrpC und TrpF. Dabei wurden für jedes Umwandlungsexperiment die Energiebeiträge der verwendeten Module für die gefundene Lösung aufgetragen. Die Module sind im Einzelnen: ROSETTA DESIGN (blau), DRUGSCORE (lila), die Funktionsdefinition (gelb) sowie PROPKA (orange). Die Energiebeitrage für das ROSETTA-Modul sind Differenzbeträge zum Testdesign des jeweiligen Gerüsts. Die Energiebeiträge für das DRUGSCORE-Modul liegen in drei Fällen außerhalb des Darstellungsbereichs: HisA auf TrpF (109), HisF auf TrpA (81) sowie HisF auf TrpF (76).

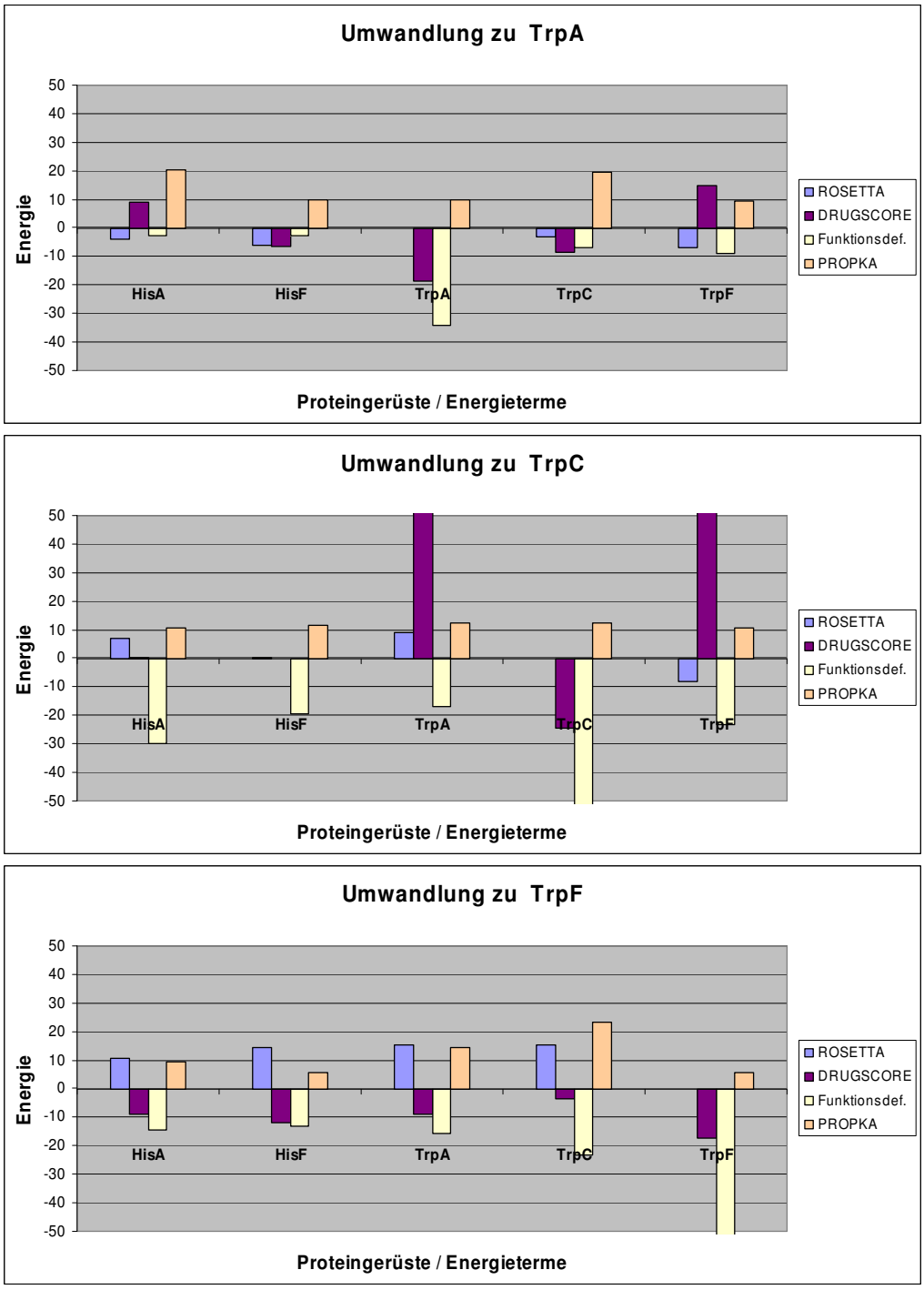


Abbildung 48: Energieprofile für TrpA, TrpC und TrpF Umwandlungen

Die aktiven Zentren von TrpA (oberes Diagramm), TrpC (mittleres Diagramm) sowie TrpF (unteres Diagramm) wurden jeweils auf fünf Proteingerüste übertragen. Dies sind HisA, HisF, TrpA, TrpC und TrpF. Dabei wurden für jedes Umwandlungsexperiment die Energiebeiträge der verwendeten Module für die gefundene Lösung als Säulenprofil aufgetragen. Die Module sind im Einzelnen: ROSETTA DESIGN (blau), DRUGSCORE (lila), die Funktionsdefinition (gelb) sowie PROPKA (orange). Die Energiebeitrage für das ROSETTA-Modul sind Differenzbeträge zum Testdesign des jeweiligen Gerüsts. Die Energiebeiträge für das DRUGSCORE-Modul liegen in zwei Fällen außerhalb des Darstellungsbereichs: TrpA auf TrpC (66) sowie TrpC auf TrpF (105). Für das Funktionsdefinitions-Modul liegt der Energiebeitrag von TrpC auf TrpC bei (-52) und von TrpF auf TrpF bei (-51).

Werden für jedes Umwandlungsexperiment die Energiebeiträge zusammengefasst und miteinander verglichen, dann lassen sich diese in drei Gruppen unterteilen (Abbildung 49). Die erste Gruppe bilden die Testdesigns mit besonders niedrigen Energiebeiträgen. Die Werte liegen zwischen 64 für TrpC und 49 für TrpA. Die zweite Gruppe bilden verschiedene Umwandlungsexperimente mit Energiewerten zwischen -25 für HisF auf HisA und +23 für TrpA auf HisA. Die dritte Gruppe bilden Umwandlungen mit deutlich höheren Energiewerten. Diese liegen zwischen +59 für HisA auf TrpA und +115 für HisA auf TrpF.

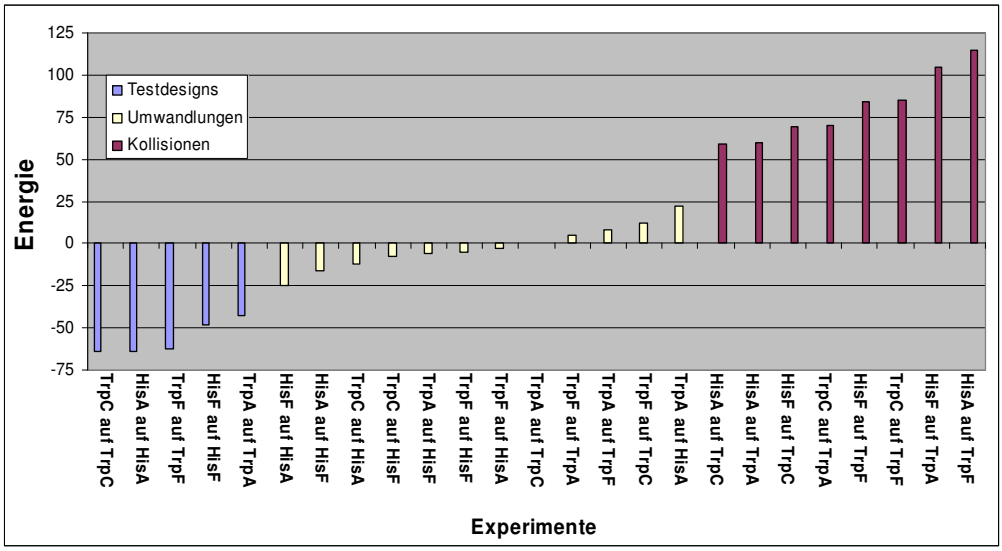


Abbildung 49: Energetische Beurteilung der Umwandlungsexperimente
Die Umwandlungsexperimente wurden hinsichtlich der Energiewerte sortiert und klassifiziert. Es wurden drei Gruppen gebildet: Umwandlungen mit niedriger Energie bestehend aus den Testdesigns (blau), Umwandlungen mit Energiewerten zwischen -25 und +25 (gelb), sowie Umwandlungen mit hoher Energie, die durch Kollisionen der Ligandenatome mit Atomen des Proteingerüsts verursacht werden (lila).

Die niedrige Energie für die Gruppe der Testdesigns resultiert aus dem idealen Gerüst für die Übertragung. Erwartungsgemäß lässt sich das aktive Zentrum jeweils auf dem wildtypischen Gerüst am besten etablieren. Die Verteilung der Energien für die zweite Gruppe zeigt, dass sich nicht alle Gerüste bzw. alle Funktionen gleich gut übertragen lassen. Um zu erklären, warum die dritte Gruppe eine so hohe Energie aufweist, müssen die einzelnen Energiebeiträge für das jeweilige Umwandlungsexperiment betrachtet werden (Abbildung 43 und Abbildung 48). In allen Fällen ist im Wesentlichen der Energiebeitrag für die Bindung verantwortlich für die hohe Gesamtenergie. Die Bindungsenergie beträgt zwischen +21 für HisF auf TrpC und +109 für HisA auf TrpF. Die hohen Werte resultieren aus Kollisionen der Ligandenatome mit dem Rückgrat des Gerüstproteins.

Dieses Problem tritt bei der Einbettung der Liganden von HisA oder HisF in den Bindetaschen von TrpA, TrpC sowie TrpF auf. Der Ligand PRFAR (HisA, HisF) ist deutlich größer als IGP (TrpA) und rCdRP (TrpF und TrpC). Damit passt PRFAR aufgrund der sterischen Hinderung nicht in das aktive Zentrum der anderen Proteingerüste (vgl. Kapitel 46). Eine erfolgreiche Modellierung dieser Umwandlungsexperimente macht eine Anpassung des Rückgrats zur Vergrößerung der Bindetasche notwendig. Da TRANSCENT auf ein starres Rückgrat beschränkt ist, lassen sich diese Umwandlungen nicht modellieren.

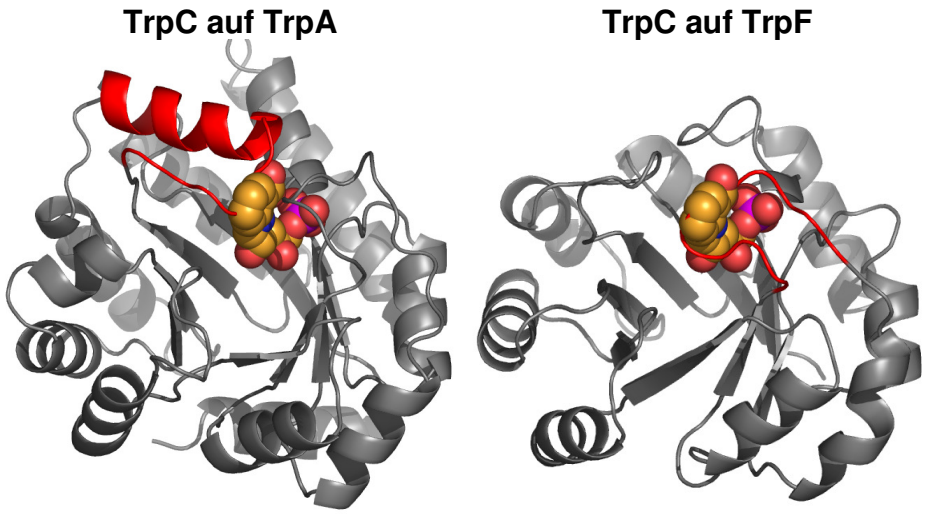


Abbildung 50: Kollisionen der Liganden von TrpC mit dem Rückgrat von TrpA und TrpF
Der erste Schritt bei der Umwandlungsmodellierung ist die Positionierung des neuen Liganden im Proteingerüst. In TRANSCENT wird der Ligand aus der Vorlage nach einer Superposition mit dem Proteingerüst übertragen. Dieser Ansatz führt bei der Umwandlung von TrpA zu TrpC (links) und TrpF zu TrpC (rechts) zu Kollisionen (rot) der Ligandenatome (Kugeldarstellung) mit den Atomen des jeweiligen Proteingerüsts (Bänderdarstellung).

Dieses Problem tritt zusätzlich bei der Umwandlung von TrpA und TrpF zu TrpC auf. Hier kollidiert der Ligand mit den Schleifenregionen des jeweiligen Proteingerüsts (Abbildung 50). Das Problem wird in diesem Fall nicht durch die Größe des Liganden, sondern von seiner Positionierung verursacht. Die Bindetasche von TrpC liegt relativ zu den Bindetaschen von TrpA und TrpF im Proteingerüst leicht versetzt. Die Liganden werden aus dem aktiven Zentrum der Vorlagestruktur übernommen. Diese werden dazu vorher mit der Struktur des Proteingerüsts superpositioniert. Diese Strategie der Ligandenpositionierung ist in diesem Fall nicht ausreichend.

4.4.3 Beispiele für Umwandlungsmodellierungen

Im Folgenden werden zwei Beispiele für besonders niedrige Energiewerte genauer betrachtet. Abbildung 49 zeigt, dass die wechselseitige Umwandlung von HisF auf HisA (Energiewert -25) sowie HisA auf HisF (Energiewert -16) besonders günstig bewertet werden.

Die Enzyme HisA und HisF sind strukturell und funktionell sehr ähnlich. Das Produkt von HisA ist das Substrat von HisF. Wildtypisches HisF aus *T. maritima* weist bereits basale HisA Aktivität auf (Lang et al., 2000). Die Sequenzidentität für HisA- und HisF-Sequenzen liegen zwischen 16% und 26% (Höcker et al., 2004). Die Sequenzen beider Proteine werden in der PFAM-Datenbank (Sonnhammer et al., 1997) unter dem gemeinsamen PFAM-Eintrag (PF00977 „*Histidine biosynthesis protein*“) verwaltet.

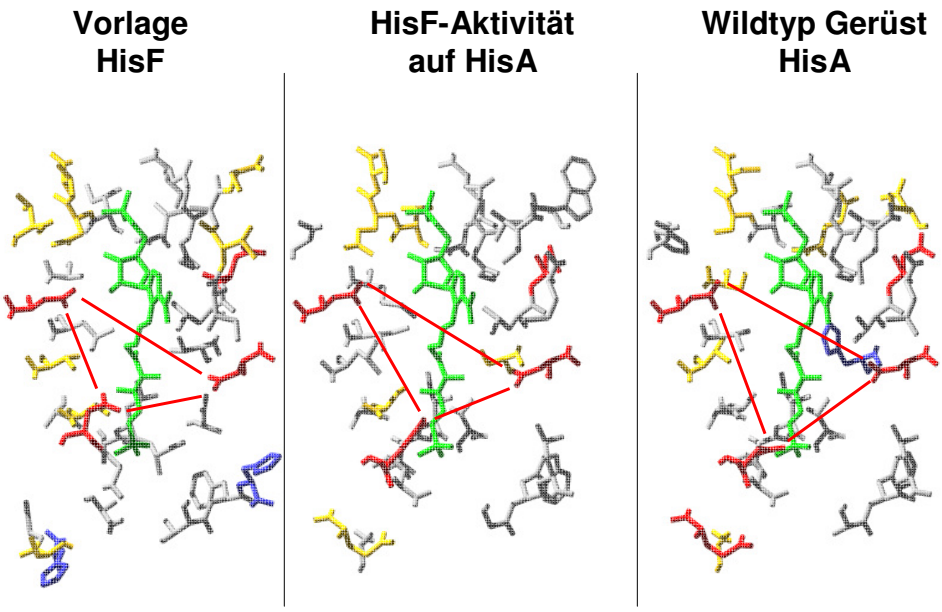


Abbildung 51: Proteingerüst von HisA mit aktivem Zentrum von HisF
Es sind die aktiven Zentren von drei verschiedenen Proteinstrukturen zu sehen. Links ist das aktive Zentrum von HisF abgebildet, das als Vorlage dient. In der Mitte ist das aktive Zentrum von HisF auf das Proteingerüst von HisA modelliert. Rechts ist das aktive Zentrum des wildtypischen HisA zu sehen, in das bereits der Ligand von HisF positioniert ist. Die roten Striche verbinden jeweils drei Aspartat-Reste an äquivalenten Positionen, die sowohl in HisF als auch in HisA in die Katalyse involviert sind. Die Abbildung zeigt Arg, His sowie Lys in blau. Asp und Glu sind rot eingefärbt. Asn, Cys, Gln, Ser, Thr sowie Tyr sind gelb gefärbt. Die hydrophoben Reste sind grau dargestellt.

Auffällig konserviert sind vor allem drei Aspartate (Abbildung 51) an jeweils äquivalenten Positionen in den beiden Enzymen (D8, D127 sowie D169 in HisA und D11, D130 sowie D176 in HisF). Die ersten beiden Aspartate sind jeweils katalytisch

essentiell und auch das jeweils dritte Aspartat hat einen Einfluss auf die katalytische Effizienz (Beismann-Driemeyer & Sterner, 2001).

Der von TRANSCENT berechnete Vorschlag für die Modellierung des aktiven Zentrums berücksichtigt diese Situation. Es werden die wildtypischen Aspartate von HisF als relevant identifiziert und in HisA an den Stellen positioniert, an denen sich auch die entsprechenden Aspartate befinden (vgl. Abbildung 51).

Im Vergleich zum Testdesign ist der vom Stabilitätsmodul berechnete Energiebeitrag leicht erhöht (+4). Daraus lässt sich ableiten, dass kein nennenswerter Unterschied in der Stabilität zu erwarten ist.

Die Bindungsenergie ist relativ niedrig (-15). Da allerdings die Bindungsenergie für das Testdesign deutlich niedriger liegt (-28), ist anzunehmen, dass sich die Einbettung des Liganden noch verbessern lässt.

Der Energiebeitrag (-31) für das Funktionsdefinitions-Modul liegt auf dem Niveau des Testdesigns (-33). Daraus lässt sich ableiten, dass die HB-Gruppenverteilung für das Modell der Vorlage sehr ähnlich ist.

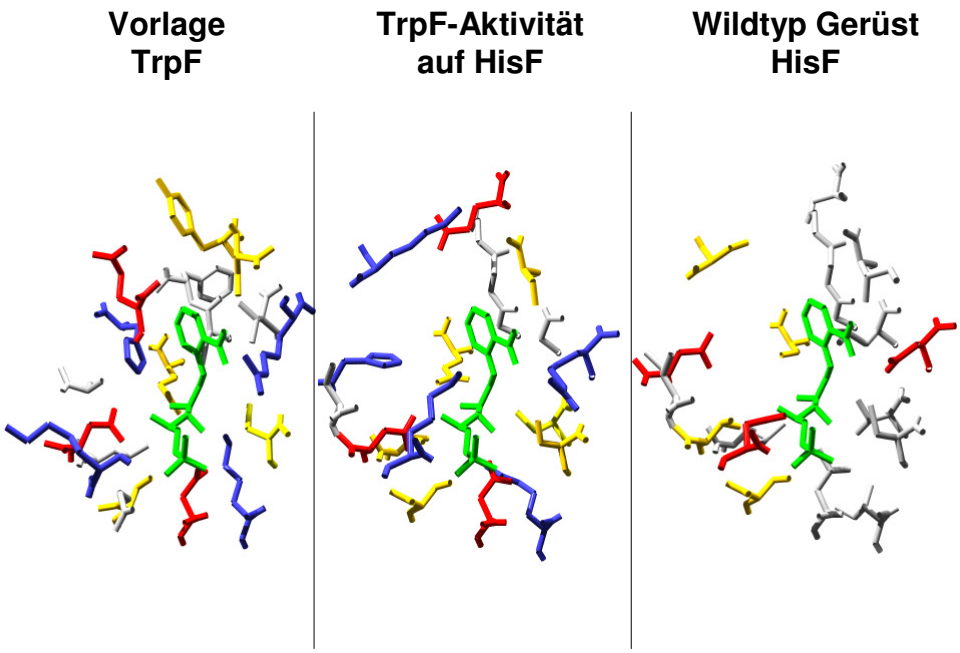


Abbildung 52: Proteingerüst von HisF mit aktivem Zentrum von TrpF
Es sind die aktiven Zentren von drei verschiedenen Proteinstrukturen zu sehen. Links ist das aktive Zentrum von TrpF abgebildet, das als Vorlage dient. In der Mitte ist das aktive Zentrum von TrpF auf das Proteingerüst von HisF modelliert. Rechts ist das aktive Zentrum des wildtypischen HisF zu sehen, in das bereits der Ligand von TrpF positioniert ist. Die Abbildung zeigt Arg, His sowie Lys in blau. Asp und Glu sind rot eingefärbt. Asn, Cys, Gln, Ser, Thr sowie Tyr sind gelb gefärbt. Die hydrophoben Reste sind grau dargestellt.

Der Energiebeitrag, der aus der pKa-Optimierung resultiert, liegt eher hoch (+18). Allerdings liegt der betreffende Wert für das Testdesign ebenfalls relativ hoch (+13). Der hohe Energiebeitrag entsteht vor allem durch ein Cystein (Position 9) und ein Histidin (Position 84) aus HisF, die auf dem Proteingerüst nicht modelliert worden sind. Da durch das Fehlen der Reste deren pKa-Wert nicht optimiert werden kann, wird ein Strafterm aufgeschlagen (vgl. Kapitel 4.1.4.2).

Das Programm modelliert aber auch solche Umwandlungen sehr plausibel, bei denen sich die aktiven Zentren von Proteingerüst und Vorlage nicht sonderlich ähneln. Dazu gehört zum Beispiel die Umwandlung von HisF zu TrpF. Das aktive Zentrum des Proteingerüsts ist wesentlich größer als das der Vorlage. Auch die Verteilung der HB-Gruppen ist in beiden aktiven Zentren sehr verschieden (vgl. Abbildung 52). So sind im aktiven Zentrum von HisF keine positiv geladenen Reste vertreten, wohingegen sich in TrpF ein Lysin (Position 5) und ein Arginin (Position 36) in unmittelbarer Nachbarschaft zum Liganden befinden.

Die Übertragung der TrpF-Aktivität auf das Gerüst von HisF wird mit einer Gesamtenergie von -5 bewertet. Der Energiebeitrag liegt für die Stabilität bei +14, für die Bindung bei -12, für die Ähnlichkeit bei -13 und für die pKa-Optimierung bei +6.

Wird das Ergebnis vergleichend mit dem Testdesign von TrpF betrachtet, so können die folgenden Schlussfolgerungen gezogen werden:

Ein Stabilitätsverlust für die Etablierung der TrpF-Aktivität ist für alle Umwandlungen gegeben (+11 für TrpF auf TrpA bis +16 für TrpF auf TrpC). Gleichzeitig wird durch Etablierung einer anderen Aktivität immer Stabilität gewonnen. (-3 für HisF auf TrpF bis -13 für HisA auf TrpF). Damit scheint der Stabilitätsverlust bei der Umwandlung von HisF zu TrpF eher in den notwendigen strukturellen Details des neuen aktiven Zentrums begründet, als in einer suboptimalen Einbettung der Reste. Das Proteingerüst stammt aus einem thermophilen Organismus, womit eine gewisse Toleranz für destabilisierende Austausche gegeben ist.

Der Energiebeitrag der Bindung (-12) ist negativ und auch der entsprechende Energiebeitrag im Testdesign ist nicht wesentlich niedriger (-17). Dies deutet darauf hin, dass der Ligand hinreichend gut in das neue aktive Zentrum eingebettet ist.

Die Ähnlichkeit der HB-Gruppen-Verteilung ist mit einem Energiebeitrag von -13 zwar gegeben, allerdings liegt der Energiebeitrag des Testdesigns mit -51 deutlich niedriger. Alle geforderten HB-Gruppen wurden zwar modelliert, ihre Verteilung entspricht jedoch nur zu einem gewissen Grad der Vorlage (vgl. Abbildung 51).

Der Energiebeitrag für den pKa-Wert liegt auf dem Niveau des Testdesigns (+6). Die einzelnen pKa-Werte entsprechen alle den Referenz-pKa-Werten (+/- 1 pH-Einheiten). Einzig der pKa-Wert für ein Cystein an Position 48 (pKa 11.8) des Modells weicht um

3.5 pH-Einheiten vom Referenz-pKa-Wert (pKa 8.3) für das Cystein an Position 7 in TrpF ab. Auch im Testdesign liegt der Wert bei 10.8, was darauf hindeutet, dass sich die Optimierung hier Vorgaben der andern Module nicht überwinden kann.

Die Aminosäuren des aktiven Zentrums der Modelle wurden durch die Optimierung neu gewählt. Auch die Seitenketten in der Umgebung (<15Å zum Liganden) des aktiven Zentrums wurden optimiert. Damit ist ein Großteil der Atome des Proteins neu modelliert worden. Um zu untersuchen, in wie weit die neuen aktiven Zentren plausibel modelliert worden sind, wurden die Modelle energieminimiert (Abbildung 53).

Für diesen Zweck wurde das Programm MOLOC (Gerber & Müller, 1995) verwendet. Das Programm ist für diesen Zweck besonders geeignet, weil es ohne aufwändige Ligandenparametrisierung auskommt. Wenn die Seitenketten verträglich modelliert sind, sollten sich die Modelle im MOLOC-Kraftfeld ähnlich verhalten wie Kristallstrukturen von Proteinen. Aus der Analyse von Kristallstrukturen mit MOLOC ist bekannt, dass energieminimierte Strukturen zur Ausgangsstruktur einen RMSD-Wert zwischen 1Å und 2Å aufweisen (Gerber & Müller, 1995). Daher kann davon ausgegangen werden, dass RMSD-Werte in dieser Größenordnung auf stabile Proteine hindeuten.

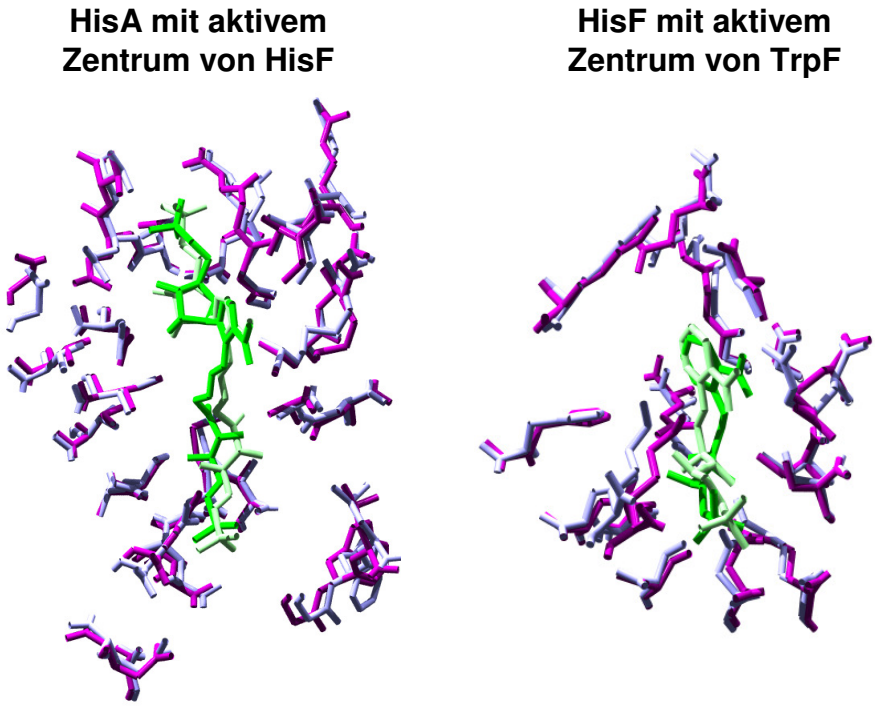


Abbildung 53: Abweichungen der Atompositionen nach Minimierung
Für die beiden Strukturmodelle der Umwandlungen von HisA mit aktivem Zentrum von HisF (links) und HisF mit aktivem Zentrum von TrpF (rechts) wurde eine Energieminimierung durchgeführt. Dazu wurde das Programm MOLOC (Gerber & Müller, 1995) verwendet. Die Atome der resultierenden Strukturen (hellblaue Strukturen mit hellgrünem Ligand) wichen von den Ausgangstrukturen (lila Strukturen mit grünem Ligand) mit einem RMSD von 0.8Å (HisA mit HisF) und 1.1Å (HisF mit TrpF) ab. Wurden nur die Atome der jeweiligen aktiven Zentren betrachtet, ergab sich ein Wert von 0.8Å sowie 1.2Å.

Für das HisA-Modell mit dem aktiven Zentrum von HisF liegt der RMSD von Modell und minimierter Struktur bei 0.8Å über alle Atome. Werden nur die Atome des aktiven Zentrums berücksichtigt, liegt der Wert ebenfalls bei 0.8Å.

Für das HisF-Modell mit dem aktiven Zentrum von TrpF liegt der RMSD zur minimierten Struktur bei 1.1Å über alle Atome. Für die Atome des aktiven Zentrums liegt der Wert bei 1.2Å.

Die Abweichungen sind mit denjenigen vergleichbar, die bei der Minimierung von Kristallstrukturen auftreten. Für Modelle, die im aktiven Zentrum grobe Fehler aufweisen wäre ein wesentlich höherer Wert zu erwarten gewesen. Somit unterstreicht dieser Befund die Plausibilität der Modelle.

An den dargestellten Umwandlungsexperimenten wurde exemplarisch gezeigt, wie sich die gefundenen Modelle auf Plausibilität überprüfen lassen. Allerdings kann eine plausible Modellierung den Erfolg einer Umwandlung nicht garantieren. Genaue Kenntnis des Modellsystems und die Analyse durch Experten können dabei helfen die Erfolgsrate zu steigern. Den Nachweis für einen Umwandlungserfolg vermag jedoch letztendlich nur das Laborexperiment liefern. Das oben skizzierte Vorgehen zeigt auf, wie mit den anderen, hier nicht detailliert vorgestellten Umwandlungsmodellierungen umgegangen werden sollte.

5 Diskussion

Im Folgenden werden einige Ergebnisse der Evaluierung nochmals aufgegriffen und im Gesamtzusammenhang diskutiert. Dabei werden auch einige Aspekte des Programmdesigns untersucht.

5.1 Das Programm TRANSCENT

In den vorherigen Kapiteln wurde der modulare Aufbau des Programms TRANSCENT beschrieben und die Leistungsfähigkeit evaluiert. In den folgenden Kapiteln werden diese Ergebnisse noch einmal in Bezug auf die Anforderungen an das Programm diskutiert.

5.1.1 Proteinstabilität

Um die Proteinstabilität zu optimieren, wird in TRANSCENT das Programm ROSETTA DESIGN verwendet. Das Programm ist für die Stabilitätsoptimierung von Proteinen entwickelt worden und nicht für den Transfer von aktiven Zentren geeignet, weil es zum Beispiel die Interaktion von Protein und Ligand nicht beschreiben kann.

Die Aminosäuren eines aktiven Zentrums sind nicht alle unmittelbar in Ligandenbindung und Katalyse involviert. Einige Aminosäuren sind vor allem für die Stabilität des Proteingerüsts relevant. ROSETTA DESIGN findet in Testdesigns etwa 30% der Aminosäuren eines aktiven Zentrums wieder. Diese sind wahrscheinlich vor allem stabilitätsrelevant, da ROSETTA für Stabilitätsoptimierung konzipiert worden ist.

Es ist anzunehmen, dass die restlichen Positionen des aktiven Zentrums weniger strukturrelevant sind. Diese Hypothese wird gestützt vom Befund, dass die zusätzlichen Module von TRANSCENT, welche funktionsrelevanten Aspekte des aktiven Zentrums beschreiben, weitere Reste in Übereinstimmung mit dem Wildtyp finden.

Der Befund, dass durch ROSETTA bei Testdesigns in aktiven Zentren etwa 30% der Reste wieder gefunden werden, stimmt gut überein mit bereits publizierten Ergebnissen (Kuhlman & Baker, 2000). Der Wert liegt zwischen 52% für Aminosäuren im Proteininneren und 27% für Aminosäuren an der Oberfläche. Aktive Zentren liegen als Bindetaschen nicht vollständig an der Proteinoberfläche, so das ein Wert >27% plausibel ist.

Die Ergebnisse aus Kapitel 4.1.1.1 zeigen, dass ROSETTA DESIGN die Region der aktiven Zentren genauer modelliert als EGAD. Da EGAD aktiv weiterentwickelt wird, besteht jedoch die Chance, dass sich zukünftige Versionen des Programms besser für

derartige Modellierung eignen. Für diesen Fall sind in TRANSCENT bereits entsprechende Schnittstellen implementiert.

Die Ergebnisse belegen, dass ein Großteil der Reste des aktiven Zentrums durch die Proteinstruktur und die Stabilitätsanforderung festgelegt sind. ROSETTA DESIGN ist in besonderem Maß für die zugedachte Funktion der Stabilitätsoptimierung in TRANSCENT geeignet, da es die damit verbundenen strukturellen Zusammenhänge hinreichend gut modellieren kann.

5.1.2 Ligandenbindung

Für die Beschreibung der Interaktion der Seitenketten des aktiven Zentrums mit dem Liganden wird in TRANSCENT das Programm DRUGSCORE verwendet. Dabei kommt eine rotamerbasierte Programmversion zum Einsatz. Diese Version erlaubt, für einzelne Rotamere die Energiebeiträge von DRUGSCORE mit denen von ROSETTA DESIGN zu verrechnen.

Allerdings kann diese Version die Lösungsmittelabschirmung der Bindetasche durch den Liganden nicht berücksichtigen. Ein derartiger Term ist zwar in DRUGSCORE vorgesehen, in der rotamerbasierten Variante aber deaktiviert, da sich auf der Basis von einzelnen Rotameren die Oberfläche der Bindetasche nicht ohne weiteres berechnen lässt. Welchen Effekt die Berücksichtigung der Desolvatation hätte, ist unklar. Anzunehmen ist, dass sich dadurch die Leistung von TRANSCENT weiter verbessern würde. Mit Verfahren um den Oberflächenanteil einzelner Rotamere abzuschätzen (Leaver-Fay et al., 2007; Pokala & Handel, 2004) könnte diese Schwachstelle behoben werden.

Eine interessante Alternative zu DRUGSCORE stellt ROSETTA LIGAND (Meiler & Baker, 2006) dar, das ebenfalls eine Energiefunktion nutzt um Protein-Ligand-Komplexe zu bewerten. Die Energiefunktion wurde explizit für die Verwendung mit ROSETTA DESIGN entwickelt. Allerdings lassen sich die Ergebnisse der Evaluierung von ROSETTA LIGAND nicht direkt mit den hier erzielten Ergebnissen vergleichen, weil zur Bewertung von ROSETTA LIGAND keine Testdesigns, sondern Dockingexperimente durchgeführt wurden. Sobald das Programm Teil des ROSETTA Softwarepakets wird, lässt sich die Energiefunktion direkt in TRANSCENT nutzen, da dann die Wechselwirkungen zwischen Rotameren und Ligand bereits durch ROSETTA DESIGN in einer Energietabelle beschrieben werden. In diesem Fall wären weder Gewichtsoptimierungen noch Anpassungen zur Integration erforderlich und die Testdesigns können unter Verwendung von ROSETTA LIGAND wiederholt werden. Somit ließe sich die Performanz beider Energiefunktionen direkt vergleichen.

Trotz der oben dargestellten Einschränkungen konnte mit der verwendeten Version die Sequenzidentität bei den Testdesigns erhöht werden. Jede Kombination von Modulen ließ sich durch die Hinzunahme des DRUGSCORE-Moduls signifikant verbessern. Die Modelle ähneln dadurch mehr dem Wildtyp, der sich durch gute Bindung des jeweiligen Liganden auszeichnet. Damit ermöglicht dieses Modul die Bindungsaffinitäten zu optimieren.

5.1.3 Die Funktionsdefinition

Um die Verteilung der HB-Gruppen in aktiven Zentren möglichst analog zur Vorlage modellieren zu können, wird in TRANSCENT für die Funktionsdefinition ein wissensbasierter Ansatz verwendet, der die Verteilung von HB-Gruppen in Beispielstrukturen erfasst.

Da die Funktionsdefinition automatisch aus Beispielstrukturen abgeleitet wird, ist für den Ableitungsprozess selbst kein Expertenwissen erforderlich. Bei der Bewertung der Häufigkeit einzelner HB-Gruppen wird angenommen, dass der Grad der Konserviertheit deren Relevanz für die enzymatische Funktion widerspiegelt. Bei der Optimierung werden Austausche bevorzugt, deren Einfluss zu einer ähnlicheren HB-Gruppenverteilung des Modells führen, wobei vor allem konservierte HB-Gruppen berücksichtigt werden. Auf diese Weise können bei der Optimierung auch gegenläufige Tendenzen differenziert gegeneinander abgewogen werden.

Dieser Ansatz bildet eine Alternative zu publizierten Verfahren, die katalytisch relevante Reste bereits vor der Optimierung starr auf dem Proteingerüst positionieren (Hellinga & Richards, 1991; Zanghellini et al., 2006). Auf diese Weise sind die HB-Gruppen der positionierten Reste fixiert und deren Verteilung über den Verlauf der Optimierung invariant. Zusätzliche HB-Gruppen, die sich im Modell ähnlich zur Vorlage platzieren lassen, können durch diesen Ansatz nicht bevorzugt berücksichtigt werden. Damit beschränkt sich im Extremfall die Ähnlichkeit zur Vorlage auf die starr fixierten HB-Gruppen.

Für die einzelnen Potentiale der Funktionsdefinition wurden relative Häufigkeitsverteilungen bestimmt, welche die räumliche Verteilung von HB-Gruppen beschreiben. Mit Hilfe dieser Häufigkeitsverteilungen wurden Wahrscheinlichkeitsdichten über eine multivariate Gaussfunktion geschätzt. Diese beschreibt die Häufigkeitsverteilung nur näherungsweise. Durch Überlagerung mehrerer Gaussfunktionen könnte die Häufigkeitsverteilung genauer angenähert werden, denn auf diese Weise würde die Feinstruktur der Häufigkeitsverteilung berücksichtigt werden. Allerdings werden mit genauer definierten Wahrscheinlichkeitsdichten die Spielräume kleiner, in denen HB-Gruppen als „*zur Verteilung ähnlich*“ klassifiziert werden können. Ein gewisser

Spielraum ist aber notwendig, um HB-Gruppen, die aufgrund von Abweichungen im Proteingerüst nicht ideal zu positionieren sind, nicht als vollständig ungeeignet zu bewerten. Um eine derartige Überanpassung („*Overfitting*“) zu vermeiden, muss die Auflösung der Wahrscheinlichkeitsdichten als Kompromiss mit dem Spielraum für die Ähnlichkeit sorgfältig eingestellt werden.

Bei gegebener Ähnlichkeit der HB-Gruppen-Verteilung des Modells zur Vorlage lässt sich dieses in Bezug auf einen Umwandlungserfolg als besonders viel versprechend einordnen. Es ist aber schwierig, mit der Ähnlichkeit allein den Umwandlungserfolg genauer abzuschätzen. In der Regel weicht ein Modell in einzelnen Details von der Vorlage ab. Mit der Ähnlichkeit lässt sich aber nicht quantitativ beurteilen welchen Einfluss solche Abweichungen auf die enzymatische Funktion haben.

Das Funktionsdefinitions-Modul kann eine differenzierte Modellierung der Ähnlichkeit automatisch durchführen. Dabei werden suboptimal modellierte Details durch entsprechende Energiebeiträge hervorgehoben. Bei der Interpretation der Abweichungen kann die automatisierte Bewertung jedoch kein Expertenwissen ersetzten.

Bei den Testdesigns steigt die Sequenzidentität durch Hinzunahme dieses Moduls deutlich. Dieses Ergebnis zeigt zum einen, dass durch den Einfluss des Moduls HB-Gruppen tatsächlich ähnlich zur Vorlage angeordnet werden. Für Testdesigns impliziert höhere Sequenzidentität eine wildtypartige Struktur. Dieses Ergebnis belegt zum anderen, dass diese Ähnlichkeit durch kein anderes Modul beschrieben wird. Die Funktionsdefinition erhält damit eine besondere Bedeutung für die Umwandlungsmodellierung, denn die Ähnlichkeit ist ein wichtiges Kriterium bei der Beurteilung der Plausibilität.

5.1.4 Optimierung der pKa-Werte

Die pKa-Wert-Optimierung wird in TRANSCENT durch die PROPKA-Methode realisiert. PROPKA bietet gute Vorhersagequalität bei vergleichsweise geringem Rechenaufwand.

Der Rechenaufwand wurde für die Verwendung in TRANSCENT weiter optimiert. Erst die etwa 5000-fache Geschwindigkeitssteigerung hat die Verwendbarkeit der Methode für den Transfer von aktiven Zentren möglich gemacht. Die Gewichtsfindung für das PROPKA-Modul dauerte trotz der Optimierung auf dem verwendeten Rechner-Cluster mehrere Tage. Mit der Geschwindigkeit der Originalversion wäre eine Gewichtsbestimmung auf diese Weise nicht möglich gewesen.

Die Genauigkeit von PROPKA wurde für die Verwendung in TRANSCENT nicht verbessert, aber durch eine konsequente Verwendung der Methode werden eventuell vorhandene systematische Ungenauigkeiten abgeschwächt. Zum einen werden mit

PROPKA die Referenz-pKa-Werte aus der Vorlagestruktur ermittelt, zum anderen wird PROPKA auch bei der Optimierung des aktiven Zentrums verwendet. Ungenau abgeschätzte pKa-Wert-Vorgaben werden dann gleichartig ungenau optimiert. Damit ist die Chance gegeben, dass sich ein solcher systematischer Fehler relativiert. Da bei der Optimierung versucht wird, das neue aktive Zentrum möglichst analog zur Vorlage zu modellieren, ist diese Annahme in besonderem Maße gerechtfertigt.

Der gebundene Ligand kann im aktiven Zentrum ebenfalls pKa-Wert-Verschiebungen verursachen. PROPKA berücksichtigt bei der pKa-Wert-Berechnung den Einfluss des Liganden jedoch nicht. Die pKa-Werte der Vorlage beschreiben daher den Protonierungszustand vor der Ligandenbindung. Entsprechend wird durch die Optimierung dieser Protonierungszustand angenähert. Der Einfluss des Liganden bleibt somit konsequent unberücksichtigt. Da die Bindungspose des Liganden aus der Vorlage übernommen wird, sollte sich dessen Einfluss gleichartig auf den Protonierungszustand von optimiertem Modell und Vorlage auswirken. Dadurch kann auch hier davon ausgegangen werden, dass sich ein eventuell nicht berücksichtigter Einfluss kompensiert.

Mit den Testdesigns konnte gezeigt werden, dass sich durch Verwendung des PROPKA-Moduls der Anteil der identischen Reste von Modellen und Vorlage erhöht. Es werden auch hier Zusammenhänge in den aktiven Zentren erfasst, die durch kein anderes Modul abgedeckt werden. Zusätzlich deutet die erhöhte Sequenzidentität an, dass sich die benötigten pKa-Wert-Verschiebungen nicht auf beliebige Weise realisieren lassen. Die gewünschten Effekte erfordern vielmehr den Einfluss von speziellen Aminosäurekonstellationen, die mit PROPKA erfasst und modelliert werden können. Da geeignet verschobene pKa-Werte eine wichtige Voraussetzung für die Katalyse darstellen und sich diese Verschiebungen offenbar nicht zwangsläufig aus der Optimierung der anderen Kriterien ergeben, wird die Bedeutung dieses Moduls für die Modellierung einer Funktionsumwandlung deutlich.

5.1.5 Kombination aller Module

Um die vier Module miteinander zu kombinieren, wurden ihre Ausgaben als Energieterme in einer gemeinsamen Energiefunktion zusammengefasst. Auf diese Weise lassen sich die Rahmenbedingungen, die durch die einzelnen Module berücksichtigt werden, gemeinsam optimieren.

Dies machte eine Gewichtung der Energiebeiträge erforderlich. Die Gewichtsoptimierung wurde separat für das DRUGSCORE-, Funktionsdefinitions- und PROPKA-Modul jeweils zusammen mit dem ROSETTA-Modul vorgenommen. Die jeweils ermittelten Gewichte wurden für die vollständige Kombination aller Module verwendet.

Es ist unklar, ob diese Gewichte auch optimal bezüglich der Gesamtkombination gewählt sind. Da die Leistungsgewinne durch die einzelnen Module nicht additiv sind, ist es möglich, dass sich Gewichte für die Gesamtkombination finden lassen, die besser geeignet sind. Für die in der vorliegenden Arbeit verwendeten Gewichte konnte aber gezeigt werden, dass sich die Effekte der einzelnen Module ergänzen. Die Gesamtkombination ist dadurch leistungsfähiger als jede Teilkombination, was die Verwendung der ermittelten Gewichte rechtfertigt.

Für die Testdesigns wird vor allem durch wildtypische Rotamere eine Konfiguration gefunden, die verträglich mit allen Rahmenbedingungen ist. Diese Tendenz spiegelt sich in den erreichten Sequenzidentitätswerten wider. Manche Reste werden in Übereinstimmung von mehreren Modulen bevorzugt. Dieser Effekt wird aus dem nicht additiven Leistungszuwachs bei Kombination der Module ersichtlich.

Tatsächlich lassen sich die einzelnen Rahmenbedingungen in aktiven Zentren nicht separat betrachten. So legt die HB-Gruppen-Verteilung einen Teil der pKa-Wert-Verschiebungen fest und HB-Gruppen vermitteln über Wasserstoffbrücken auch Ligandenbindung. In solchen Fällen überlagern sich die Präferenzen der einzelnen Module. Überlagerte Präferenzen spiegeln also zu einem gewissen Grad die Bedeutung der jeweiligen Reste wider.

Im Anwendungsfall, d.h. wenn ein aktives Zentrum zwischen zwei verschiedenen Proteinen übertragen wird, weist das Proteingerüst in der Regel strukturelle Unterschiede zur Vorlage auf. Dann lassen sich nicht immer Reste finden, die alle Rahmenbedingungen optimal erfüllen. In solchen Fällen erhöhen überlagerte Präferenzen die Wahrscheinlichkeit, dass geeignete Reste trotzdem zuverlässig gewählt werden.

Gegenüber dem ROSETTA-Modul ist der Einfluss der anderen Module durch die Gewichtung eher schwach eingestellt. Bei Testdesigns sind diese Module in der Lage, Präferenzen des ROSETTA-Moduls für nicht wildtypische Aminosäuren zu kompensieren. Die dadurch zusätzlich gewählten wildtypischen Aminosäuren stellen in der Regel keinen großen Kompromiss für die Stabilität dar, da sie auch auf dem Gerüst des Wildtyps vorkommen. Dieser Kompromiss lässt sich durch die Energiedifferenz zwischen der optimal stabilen Sequenz, die durch ROSETTA DESIGN allein gewählt wird und der wildtypartigen Sequenz, die in Kombination mit den jeweiligen Modulen gewählt wird, ausdrücken. Die maximal überwindbare Energiedifferenz bildet den Einflussspielraum, der den zusätzlichen Modulen im Anwendungsfall zur Verfügung steht.

Es kann daher sinnvoll sein, die Gewichtung im Anwendungsfall anzupassen. Es bietet sich an, die entsprechenden Gewichte zu erhöhen, wenn abgeschätzt werden soll bis zu welchem Grad eine Rahmenbedingung maximal erfüllbar ist. Wird beispielsweise die

Gewichtung für das PROPKA-Modul erhöht, so wird das Optimierungsverfahren vor allem pKa-Werte optimieren und dabei die anderen Rahmenbedingungen vernachlässigen. Ist eine Rahmenbedingung trotz hohem Gewicht nicht erfüllbar, lässt sie sich auch in keiner anderen Gewichtungskombination erfüllen. Durch Modulation der Gewichte lässt sich somit für suboptimale Modellierungen entscheiden, ob diese aus dem Kompromiss zwischen den einzelnen Rahmenbedingungen resultieren oder ob sie eine grundsätzliche Limitation des verwendeten Proteingerüsts darstellen.

Ist auf Grund von Expertenwissen eine geeignete Alternative zu suboptimalen Modellierungen bekannt, so kann diese durch einen manuellen Eingriff in den Optimierungsprozess erzwungen werden. Dies wird durch eine entsprechende Gewichtung der dazu notwendigen Aminosäuren oder Rotameren erreicht.

Mit TRANSCENT ist ein vollständiger Prozess zur Modellierung der Funktionsübertragung zwischen Enzymen realisiert worden. Die Bewertung der Gesamtkombination belegt, dass die einzelnen Module auch bei einer gemeinsamen Verwendung auf geeignete Weise die Einhaltung und Optimierung der jeweiligen Rahmenbedingungen sicherstellen. Mit Hilfe der Energiebeiträge zur gefundenen Lösung lässt sich analysieren, zu welchem Grad die einzelnen Rahmenbedingungen erfüllt worden sind. Durch die Kombination der verwendeten Methoden stellt TRANSCENT eine vielversprechende Alternative zu bestehenden Ansätzen dar. Vor allem die Optimierung der Ähnlichkeit und der pKa-Werte eröffnet dabei neue Möglichkeiten bei Umwandlungsmodellierungen.

5.2 Strukturbibliothek und Homologiemodelle

Für die automatisierte Ableitung der Funktionsdefinition benötigt TRANSCENT eine Strukturbibliothek. Die Zusammenstellung und Qualität der darin enthalten Strukturen beeinflusst den Modellierungsprozess. Daher wird im Folgenden auf diese beiden Aspekte genauer eingegangen.

5.2.1 Zusammensetzung der Strukturbibliothek

Da die Menge bekannter Strukturen zu einem Protein im Allgemeinen nicht ausreicht, um die strukturelle Variabilität des zu modellierenden aktiven Zentrums zu erfassen, wird die Menge der Strukturen durch Homologiemodelle ergänzt. Die genaue Zusammensetzung der Strukturbibliothek ist für die Arbeitsweise von TRANSCENT aber unerheblich. Daher ist es möglich und wahrscheinlich auch sinnvoll, anderweitige Vorlagestrukturen zu verwenden. So könnten auch andere Homologiemodellierungsprogramme genutzt werden um Strukturmodelle zu generieren (Ginalski, 2006). Eine

gemeinsame Verwendung mehrerer Programme würde es erlauben, spezifische Schwächen zu kompensieren.

Auch für die PFAM-Datenbank, die in der vorliegenden Arbeit als Datenquelle für homologe Sequenzen genutzt wurde, sind Alternativen denkbar.

Der Testdatensatz musste von 128 Proteinen auf 27 reduziert werden, weil für die restlichen Proteine entweder kein PFAM-Eintrag existierte oder die Anzahl der homologen Sequenzen zu gering war. Daraus wird ersichtlich, dass es im Einzelfall sogar notwendig sein kann, alternative Datenquellen zu berücksichtigen. Die einzige Bedingung für die Auswahl der Sequenzen ist, dass sie von Homologen des Vorlageenzyms stammen, da alle modellierten Strukturmodelle Positivbeispiele für sie zu übertragende Funktion darstellen müssen.

Ein besonders vielversprechender Ansatz, die Strukturbibliothek zu erweitern, wäre die Verwendung von NMR-Strukturbündel (Wüthrich, 1995). Diese sind zum einen qualitativ hochwertig, zum anderen beschreiben sie die strukturelle Variabilität einzelner Vorlageproteine. Da aber für Proteine selten mehr als eine aufgeklärte NMR-Struktur existiert, ist eine alleinige Verwendung von NMR-Strukturbündeln nicht möglich. Die Strukturbibliothek lässt sich aber auch mit einzelnen Strukturbündeln sinnvoll ergänzen.

5.2.2 Stichprobenumfang und Normierung

Die Beispielstrukturen in der Strukturbibliothek bilden eine Strichprobe von Positivbeispielen für ein aktives Zentrum mit spezifischer katalytischer Funktion. Aus den Gemeinsamkeiten und Unterschieden der Stichprobe wird für TRANSCENT die Funktionsdefinition abgeleitet.

Damit die Wahrscheinlichkeitsverteilung zuverlässig geschätzt werden können, sollte die Stichprobe möglichst umfangreich sein. Um eine Verzerrung durch überrepräsentierte Beispiele zu vermeiden, sollte die Stichprobe außerdem bereinigt werden.

Die notwendige Stichprobengröße wird bei den Strukturbeispielen der Strukturbibliothek erreicht, indem die Menge der bekannten Strukturen um Homologiemodelle erweitert wird.

Die Bereinigung der Strukturbibliothek von überrepräsentierten Beispielen ist dagegen schwierig. Üblicherweise wird eine Menge von Sequenzen, die als Stichprobe verwendet werden soll, dadurch bereinigt, dass sehr ähnliche Sequenzen gefiltert werden.

Die Homologiemodelle werden aus Sequenzen generiert, indem sie auf eine Vorlagestruktur, dem Templat, modelliert werden. Da die Modellierungsqualität der Homologiemodelle entscheidend von der Ähnlichkeit der Sequenz zum Templat

abhängt, werden nur ähnliche Sequenzen berücksichtigt. Varianten des aktiven Zentrums, die in ihrer Sequenz stark vom aktiven Zentrum des Templats abweichen, aber trotzdem die Funktion realisieren, werden ignoriert.

Damit sind Strukturmodelle in der Strukturbibliothek überrepräsentiert, die dem Templat ähnlich sind. Also kann bei der Optimierung nicht die ganze Variabilität genutzt werden, die dem aktiven Zentrum unter Erhalt der Funktion theoretisch zur Verfügung stehen würde.

5.2.3 Schwellen

Um die Qualität der Homologiemodellierungen sicher zu stellen, wurden Testmodellierungen mit einem Datensatz von Ribulosephosphat-$(\beta\alpha)_8$-Barrel-Enzymen durchgeführt. Hier zeigte sich, dass vor allem das Alignment von Eingabe-Sequenz und Sequenz der Templatstruktur für die Modellierungsqualität verantwortlich ist. Falsch zugeordnete Positionen führten zu extremen Fehlern in den modellierten aktiven Zentren. Aber es konnten Schwellen für die Sequenzähnlichkeit und die Alignmentqualität bestimmt werden, so dass für den Testdatensatz sichergestellt werden konnte, dass der Modellierungsfehler innerhalb eines Toleranzbereichs blieb.

Die Schwellen wurden für einen Testdatensatz bestimmt, der nur aus Ribulosephosphat-$(\beta\alpha)_8$-Barrel-Enzymen bestand. Daher lassen sich die aus der Analyse gewonnenen Erkenntnisse nicht uneingeschränkt verallgemeinern.

Für die Modellierung anderer Enzymklassen können in einem ersten Ansatz die hier bestimmten Schwellen verwendet werden. Sicherlich ist es jedoch sinnvoller, die in der vorliegenden Arbeit beschriebenen Testmodellierungen für andere Enzymklassen zu wiederholen, um so auch in diesen Fällen eine konkrete Modellierungsqualität sicherstellen zu können.

Die verwendeten Schwellen wurden speziell für die Anwendung in TRANSCENT ermittelt, sie stellen aber auch über den Kontext der vorliegenden Arbeit hinaus einen Erkenntnisgewinn dar. Werden für anderweitige Zwecke Homologiemodelle von Ribulosephosphat-$(\beta\alpha)_8$-Barrel-Enzymen auf die beschriebene Weise generiert, lässt sich auch in diesen Fällen beurteilen, zu welchem Grad den Strukturdetails in den aktiven Zentren der Modelle vertraut werden kann.

6 Ausblick

In der Diskussion sind bereits Verbesserungsmöglichkeiten zu einzelnen Implementierungsdetails besprochen worden. In diesem Kapitel wird ausgeführt durch welche Zusatzfunktionen die Leistungsfähigkeit weiter verbessert und das Anwendungsspektrum vergrößert werden könnte.

6.1 Ligandenpositionierung

Da mit TRANSCENT nur aktive Zentren zwischen Proteinen mit gleicher Faltung getauscht werden sollen, wurde auf eine aufwändige Ligandenpositionierung verzichtet. Der Ligand wird für die Modellierung aus der Vorlage übernommen, nachdem die Vorlagestruktur mit dem Proteingerüst des Modells superpositioniert worden ist. Dadurch wird er an einer zur Vorlage äquivalenten Stelle im Modell positioniert.

In einigen Fällen war diese Vorgehensweise nicht ausreichend, da die Atome des Liganden mit den Atomen des Proteingerüsts kollidierten. In den meisten Fällen ließ sich dieses Problem auf eine zu kleine Bindetasche (HisA oder HisF auf TrpA, TrpF oder TrpC) im Proteingerüst zurückführen. In den Fällen „*TrpC auf TrpF*“ und „*TrpC auf TrpA*“ jedoch wurde der Ligand durch die Superpositionierung ungeeignet positioniert.

Es wäre daher sinnvoll, das Programm um eine Ligandenpositionierungsroutine zu ergänzen, wie sie z.B. in (Hellinga & Richards, 1991; Zanghellini et al., 2006) beschrieben ist. Damit kann der Ligand optimal im Proteingerüst platziert werden.

Aus der flexiblen Ligandenpositionierung ergeben sich zwei weitere Vorteile: Zum einen würde die Funktionsdefinition um die explizite Positionierung der katalytisch essentiellen Resten ergänzt. Potentielle Bindetaschen im Proteingerüst werden nicht nur danach beurteilt, ob sie den Liganden aufnehmen können, sondern auch, ob sich relativ zum Liganden die essentiellen Reste in Katalyse-relevanter Ausrichtung in der Bindetasche positionieren lassen. Damit wäre schon vor der eigentlichen Optimierung des aktiven Zentrums sicher gestellt, dass sich die katalytisch essentiellen Reste prinzipiell geeignet positionieren lassen.

Zum anderen wäre TRANSCENT nicht mehr an die Bedingung des gemeinsamen Faltungsmusters von Vorlage und Proteinmodell gebunden. Aktive Zentren könnten dann auch zwischen Proteinen mit unterschiedlicher Faltung übertragen werden, wie es z.B. mit DEZYMER (Hellinga & Richards, 1991) möglich ist.

6.2 Weitere Module

TRANSCENT deckt mit den vorgestellten Modulen vier wichtige Rahmenbedingungen bei der Umwandlungsmodellierung ab. Es ist aber möglich, dass die Berücksichtigung zusätzlicher Rahmenbedingungen die Modellierungsgenauigkeit weiter erhöhen würde.

Für bestimmte katalytische Prozesse wird angenommen, dass die Substratbindung mit einer Schleifenbewegung einhergeht, die das aktive Zentrum verschließt und so das Substrat vom Lösungsmittel abschirmt (Xiang et al., 2004). Ist eine derartige Bewegung für den Reaktionsverlauf notwendig, kann es für den Umwandlungserfolg entscheidend sein, dass die notwendige Flexibilität der entsprechenden Region bei der Modellierung berücksichtigt wird.

Die breite funktionale Vielfalt der Enzyme mit $(\beta\alpha)_8$-Barrel-Faltung wird unter anderem durch hohe Variabilität in den Schleifenregionen des Katalysepols erreicht. Dabei variieren einzelnen Schleifen nicht nur in ihrer Sequenz, sondern auch in ihrer Länge (Sterner & Höcker, 2005). Auf diese Weise entstehen in den Schleifenregionen zum Teil sehr unterschiedliche Topologien. Die Berücksichtigung alternativer Schleifentopologien könnte strukturelle Einschränkungen eines Proteingerüsts, wie etwa eine zu kleine Bindetasche, kompensieren.

Neben struktureller Stabilität ist vor allem die Löslichkeit eine weitere wichtige Eigenschaft von Proteinen. Wenn Modifikationen, wie die Umwandlung des aktiven Zentrums zu einem Verlust der Löslichkeitseigenschaften führen, lässt sich der Umwandlungserfolg nicht mehr beurteilen (Riede, 2006).

Es wäre daher sinnvoll, auch die Löslichkeit eines Proteins bei der Modellierung zu bewerten und zu optimieren. So ließen sich die Erfolgsaussichten bei Umwandlungsexperimenten weiter steigern.

Erweiterungen dieser Art würden das Programm effizienter und universeller verwendbar machen. Die Berücksichtigung dieser Rahmenbedingungen stellt aber eine enorme technische Herausforderung dar, da hierfür noch keine Ansätze vorgeschlagen und umgesetzt wurden, die sich unmittelbar integrieren ließen.

6.3 *De Novo* Funktionsdefinitionen

Funktionsdefinitionen werden in dieser Arbeit immer aus aktiven Zentren mit bereits bekannter Proteinfunktion abgeleitet. Dies bedeutet, dass der Modellierungsprozess eine Funktionstransplantation vornimmt, also eine bereits bekannte Funktion im Proteinmodell nachbildet. Im Prinzip ist es auch möglich eine neue Proteinfunktion zu „*erfinden*“ und diese dann in einer Funktionsdefinition zu beschreiben. Allerdings sind

nicht alle biochemischen Vorgänge in ausreichendem Umfang verstanden, was diesen Ansatz außerordentlich schwierig macht.

Bereits für Proteine, deren Katalysemechanismus und Struktur bekannt ist, stellt die Beschreibung der Funktionsdefinition ein nicht triviales Problem dar. Entsprechend schwieriger ist es, eine Funktionsdefinition festzulegen, wenn der Reaktionsablauf unklar ist oder wenn keine Struktur bekannt ist. Das „Erfinden“ neuer Funktionsdefinitionen wäre nochmals eine Stufe komplizierter. In der vorliegenden Arbeit wurde diese Problematik umgangen, in dem die Funktionsdefinition wissensbasiert aus der Strukturbibliothek abgeleitet worden ist.

Konzeptionell ist TRANSCENT aber nicht auf die Verwendung abgeleiteter Funktionsdefinitionen beschränkt. Eine anderweitig generierte Funktionsdefiniton könnte ebenfalls in die Optimierung einfließen. Wenn es also in Zukunft gelingt, neue Funktionsdefinitionen zu „*erfinden*“, könnten auch solche die Grundlage für die Modellierung bilden. Somit könnte TRANSCENT dazu dienen *De Novo* Enzyme zu entwickeln.

Danksagung

Dieses Kapitel ist den Leuten gewidmet, ohne die das Gelingen dieser Arbeit nicht möglich gewesen wäre.

Ich danke meinem Doktorvater Dr. PD Rainer Merkl für eine wunderbare Betreuung. Er stand mir immer mit Rat und Tat zur Seite und ermutigte mich, diese Arbeit nach meinen eigenen Vorstellungen zu gestalten. Mit seiner langjährigen Erfahrung in der Bioinformatik, half er mir Probleme und Ergebnisse richtig einzuschätzen.

Großer Dank gebührt Herrn Prof. Dr. Sterner, dessen Forschungsinteressen die Basis dieser Arbeit begründet haben. Durch seine Vision, Enzymevolution auch mit bioinformatischen Methoden zu erforschen, habe ich als Informatiker meine Aufgabe in der Biochemie gefunden.

Herrn Dr. PD Wolfram Gronwald danke ich für die Übernahme der Funktion als Zweitgutachter und für ergiebige Diskussionen und Hilfestellungen rund um das Thema Strukturbiologie.

Dr. Marco Bocola danke ich für seine Unterstützung und für viele entscheidende Impulse zu dieser Arbeit. Mit ihm konnte ich Modellierungsprobleme sowohl biochemischer als auch technischer Art diskutieren.

Ich möchte Herrn Prof. Dr. Stefan Dove für seine Hilfsbereitschaft und für aufschlussreiche Diskussionen danken. Durch ihn wurde mir die Ähnlichkeit vieler Fragestellungen des Wirkstoffdesigns und des Proteindesigns klar.

Allen Sternern danke ich für geduldiges Erklären und Diskutieren biochemischer Zusammenhänge, für viele interessante Einblicke in die verschiedenen Projekte und für eine gute Lehrstuhlatmosphäre.

Meinem heiteren Zimmerkollegen Hermann Zellner danke ich, dass er mir die bayrische Lebensart und die Vorzüge von Linux näher brachte.

Ich danke Felix List für viele einleuchtende Erklärungen, die mein Biochemie-verständnis verbessert haben. Gerne erinnere ich mich an unterhaltsame Diskussionen, in denen er mir die Unzulänglichkeit manch grenzwissenschaftlich anmutender Idee aufzeigte.

Dr. Helmut Durchschlag danke ich für viele interessante Gespräche über Wissenschaft und Forschung und Jörg Claren für solidarische Nacht- und Wochenendschichten.

Ich danke Markus Richter für sein Organisationstalent und Alexander Ehrmann für kameradschaftliches Teilen von Zimmern und Fischen. Beiden danke ich außerdem für interessante fachliche Diskussionen.

Klaus Tiefenbach danke ich für seine Unterstützung bei der Hardwarebeschaffung und bei den alltäglichen Netzwerkproblemen und Matthias Zwick für Hilfestellungen rund um das Thema MSA.

Ich danke den Praktikanten und Diplomanden der Bioinformatik Marion Strieder, Helga Wagner, Nils Enkler, Martin Erlekamm, Korbinian Stöckl, Martin Ostermaier und Ariane Felgenträger für ihre Beiträge und für eine kurzweilige Zeit.

Außerdem danke ich Gerd Neudert für eine Spezialversion des Programms DRUGSCORE.

Meinem Studienfreund Tim Pohle möchte ich für seine Hilfsbereitschaft und Kompetenz bei technischen Fragestellungen danken und für unterhaltsame Besuche beim Pendeln zwischen Linz (a.d.D.) und Linz (a.R.).

Meinem Vater und meinen Großeltern danke ich für ihre Unterstützung und ihr Vertrauen in mich. Auf sie konnte ich mich immer verlassen.

Besonders möchte ich mich bei meiner Frau Ana und meinem Sohn André für ihre Liebe und Unterstützung bedanken. Sie hatten immer Verständnis, wenn es an der Uni viel zu tun gab und heiterten mich dann mit spontanen Kurzbesuchen auf.

Schließlich möchte ich allen für eine gute gemeinsame Zeit danken, die mir viel Spaß bereitet und viele interessante Einblicke gewährt hat.

Literaturverzeichnis

Alifano, P., Fani, R., Lio, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M. S. & Bruni, C. B. (1996). Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev **60:** 44-69

Allen, B. D. & Mayo, S. L. (2006). Dramatic performance enhancements for the FASTER optimization algorithm. J Comput Chem **27:** 1071-1075

Allen, F. H. (2002). The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr B **58:** 380-388

Allert, M., Dwyer, M. A. & Hellinga, H. W. (2007). Local encoding of computationally designed enzyme activity. J Mol Biol **366:** 945-953

Allert, M., Rizk, S. S., Looger, L. L. & Hellinga, H. W. (2004). Computational design of receptors for an organophosphate surrogate of the nerve agent soman. Proc Natl Acad Sci U S A **101:** 7907-7912

Altamirano, M. M., Blackburn, J. M., Aguayo, C. & Fersht, A. R. (2000). Directed evolution of new catalytic activity using the $(\beta\alpha)_8$-barrel scaffold. Nature **403:** 617-622

Altamirano, M. M., Blackburn, J. M., Aguayo, C. & Fersht, A. R. (2002). Retraction. Directed evolution of new catalytic activity using the $(\beta\alpha)_8$-barrel scaffold. Nature **417:** 468

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D. & Waley, S. G. (1975). Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Å resolution using amino acid sequence data. Nature **255:** 609-614

Barona-Gomez, F. & Hodgson, D. A. (2003). Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis. EMBO Rep **4:** 296-300

Barth, P., Alber, T. & Harbury, P. B. (2007). Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. Proc Natl Acad Sci U S A **104:** 4898-4903

Bashford, D. & Karplus, M. (1990). pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry **29:** 10219-10225

Beismann-Driemeyer, S. & Sterner, R. (2001). Imidazole glycerol phosphate synthase from Thermotoga maritima. Quaternary structure, steady-state kinetics, and reaction mechanism of the bienzyme complex. J Biol Chem **276:** 20387-20396

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol **112:** 535-542

Boas, F. E. & Harbury, P. B. (2007). Potential energy functions for protein design. Curr Opin Struct Biol **17:** 199-204

Böhm, H.-J., Klebe, G. & Kubinyi, H. (1996). Wirkstoffdesign, Heidelberg, Berlin, Oxford

Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. Proc Natl Acad Sci U S A **98:** 14274-14279

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science **253:** 164-170

Caetano-Anolles, G., Kim, H. S. & Mittenthal, J. E. (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proc Natl Acad Sci U S A **104:** 9358-9363

Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci **12:** 2001-2014

Carugo, O. & Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci **10:** 1470-1473

Charlebois, R. L., Sensen, C. W., Doolittle, W. F. & Brown, J. R. (1997). Evolutionary analysis of the hisCGABdFDEHI gene cluster from the archaeon *Sulfolobus solfataricus* P2. J Bacteriol **179:** 4429-4432

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. Embo J **5:** 823-826

Chowdry, A. B., Reynolds, K. A., Hanes, M. S., Voorhies, M., Pokala, N. & Handel, T. M. (2007). An object-oriented library for computational protein design. J Comput Chem **28:** 2378-2388

Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. Science **278:** 82-87

Dantas, G., Corrent, C., Reichow, S. L., Havranek, J. J., Eletr, Z. M., Isern, N. G., Kuhlman, B., Varani, G., Merritt, E. A. & Baker, D. (2007). High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. J Mol Biol **366:** 1209-1221

Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol **332:** 449-460

Darimont, B., Stehlin, C., Szadkowski, H. & Kirschner, K. (1998). Mutational analysis of the active site of indoleglycerol phosphate synthase from *Escherichia coli*. Protein Sci **7:** 1221-1232

Davies, M. N., Toseland, C. P., Moss, D. S. & Flower, D. R. (2006). Benchmarking pKa prediction. BMC Biochem **7:** 18

de Lorimier, R. M., Smith, J. J., Dwyer, M. A., Looger, L. L., Sali, K. M., Paavola, C. D., Rizk, S. S., Sadigov, S., Conrad, D. W., Loew, L. & Hellinga, H. W. (2002). Construction of a fluorescent biosensor family. Protein Sci **11:** 2655-2675

De Maeyer, M., Desmet, J. & Lasters, I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. Fold Des **2:** 53-66

DeLano, W. L. (2002). The PyMOL Molecular Graphics System. DeLano Scientific Palo Alto, CA, USA**:**

Desjarlais, J. R. & Clarke, N. D. (1998). Computer search algorithms in protein modification and design. Curr Opin Struct Biol **8:** 471-475

Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. J Mol Biol **290:** 305-318

Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. Nature **356:** 539 - 542

Desmet, J., Spriet, J. & Lasters, I. (2002). Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. Proteins **48:** 31-43

Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D. & Voelz, V. A. (2007). The protein folding problem: when will it be solved? Curr Opin Struct Biol**:**

Dunbrack, R. L., Jr. (2002). Rotamer libraries in the 21st century. Curr Opin Struct Biol **12:** 431-440

Dunbrack, R. L., Jr. & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci **6:** 1661-1681

Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol **230:** 543-574

Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2004). Computational design of a biologically active enzyme. Science **304:** 1967-1971

Eijsink, V. G., Bjork, A., Gaseidnes, S., Sirevag, R., Synstad, B., van den Burg, B. & Vriend, G. (2004). Rational engineering of enzyme stability. J Biotechnol **113:** 105-120

Enkler, N. (2006). Entwicklung und Validierung einer Schnittstelle zum Modellierungs- und Energieberechnungsmodul von Rosetta Design. Diplomarbeit**:**

Eswar, N., Eramian, D., Webb, B., Min-Shen, Y. & Sali, A. (2007). Protein Structure Modeling With MODELLER. Tutorial from wwwsalilaborg**:**

Fogolari, F., Brigo, A. & Molinari, H. (2002). The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. J Mol Recognit **15:** 377-392

Georgiev, I. & Donald, B. R. (2007). Dead-end elimination with backbone flexibility. Bioinformatics **23:** i185-194

Gerber, P. R. & Müller, K. (1995). MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. J Comput Aided Mol Des **9:** 251-268

Gerlt, J. A. & Babbitt, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem **70:** 209-246

Gherardini, P. F., Wass, M. N., Helmer-Citterich, M. & Sternberg, M. J. (2007). Convergent evolution of enzyme active sites is not a rare phenomenon. J Mol Biol **372:** 817-845

Ginalski, K. (2006). Comparative modeling for protein structure prediction. Curr Opin Struct Biol **16:** 172-177

Glasner, M. E., Gerlt, J. A. & Babbitt, P. C. (2007). Mechanisms of protein evolution and their application to protein engineering. Adv Enzymol Relat Areas Mol Biol **75:** 193-239, xii-xiii

Gohlke, H., Hendlich, M. & Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol **295:** 337-356

Gohlke, H. & Klebe, G. (2002). Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed Engl **41:** 2644-2676

Goldstein, R. F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J **66:** 1335-1340

Gordon, D. B., Hom, G. K., Mayo, S. L. & Pierce, N. A. (2003). Exact rotamer optimization for protein design. J Comput Chem **24:** 232-243

Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. Curr Opin Struct Biol **9:** 509-513

Gordon, D. B. & Mayo, S. L. (1999). Branch-and-terminate: a combinatorial optimization algorithm for protein design. Structure **7:** 1089-1098

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A **84:** 4355-4358

Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol **320:** 369-387

Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis **18:** 2714-2723

Hellinga, H. W. & Richards, F. M. (1991). Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. J Mol Biol **222:** 763-785

Hendlich, M. (1998). Databases for protein-ligand complexes. Acta Crystallogr D Biol Crystallogr **54:** 1178-1182

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A **89:** 10915-10919

Henn-Sax, M., Höcker, B., Wilmanns, M. & Sterner, R. (2001). Divergent evolution of $(\beta\alpha)_8$-barrel enzymes. Biol Chem **382:** 1315-1320

Henn-Sax, M., Thoma, R., Schmidt, S., Hennig, M., Kirschner, K. & Sterner, R. (2002). Two $(\beta\alpha)_8$-barrel enzymes of histidine and tryptophan biosynthesis have similar reaction mechanisms and common strategies for protecting their labile substrates. Biochemistry **41:** 12032-12042

Hirayama, N. (2007). Docking method for drug discovery. Yakugaku Zasshi **127:** 113-122

Höcker, B., Claren, J. & Sterner, R. (2004). Mimicking enzyme evolution by generating new (betaalpha)8-barrels from (betaalpha)4-half-barrels. Proc Natl Acad Sci U S A **101:** 16448-16453

Holland, J. (1993). Adaptation in Natural and Artificial Systems, Cambridge, Mass.

Jaenicke, R. (2000). Stability and stabilization of globular proteins in solution. J Biotechnol **79:** 193-203

Johannes, T. W. & Zhao, H. (2006). Directed evolution of enzymes and biosynthetic pathways. Curr Opin Microbiol **9:** 261-267

Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J Am Chem Soc **118:** 11225-11236

Jürgens, C., Strom, A., Wegener, D., Hettwer, S., Wilmanns, M. & Sterner, R. (2000). Directed evolution of a $(\beta\alpha)_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways. Proc Natl Acad Sci U S A **97:** 9925-9930

Karlin, S. & Brocchieri, L. (1996). Evolutionary conservation of RecA genes in relation to protein structure and function. J Bacteriol **178:** 1881-1894

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30:** 3059-3066

Khersonsky, O., Roodveldt, C. & Tawfik, D. S. (2006). Enzyme promiscuity: evolutionary and mechanistic aspects. Curr Opin Chem Biol **10:** 498-508

Kirkpatrick, S., Gelatt, C. & Vecchi, M. (1983). Optimization by simulated annealing. Science **220,4598:** 671--680

Koehl, P. & Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J Mol Biol **239:** 249-275

Kohlbacher, O. & Lenhof, H. P. (2000). BALL--rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library. Bioinformatics **16:** 815-824

Krieger, E., Nielsen, J. E., Spronk, C. A. & Vriend, G. (2006). Fast empirical pKa prediction by Ewald summation. J Mol Graph Model **25:** 481-486

Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A **97:** 10383-10388

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science **302:** 1364-1368

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. Naval Research Logistics**:**

Kuper, J., Dönges, C. & Wilmanns, M. (2005). Two-fold repeated $(\beta\alpha)_4$ half-barrels may provide a molecular tool for dual substrate specificity. EMBO Rep **6:** 134-139

Lambeck, I. (2004). Arbeiten zur Umwandlung der Tryptophansynthase α-Untereinheit aus *Thermotoga maritima* in eine Triosephosphatisomerase. Diplomarbeit**:**

Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. (2000). Structural evidence for evolution of the $(\beta\alpha)_8$-barrel scaffold by gene duplication and fusion. Science **289:** 1546-1550

Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. (2006). Combinatorial methods for small-molecule placement in computational enzyme design. Proc Natl Acad Sci U S A **103:** 16710-16715

Lasters, I., De Maeyer, M. & Desmet, J. (1995). Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. Protein Eng **8:** 815-822

Leaver-Fay, A., Butterfoss, G. L., Snoeyink, J. & Kuhlman, B. (2007). Maintaining solvent accessible surface area under rotamer substitution for protein design. J Comput Chem **28:** 1336-1341

Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. J Mol Biol **236:** 918-939

Lee, C. & Subbiah, S. (1991). Prediction of protein side-chain conformation by packing optimization. J Mol Biol **217:** 373-388

Leisola, M. & Turunen, O. (2007). Protein engineering: opportunities and challenges. Appl Microbiol Biotechnol **75:** 1225-1232

Leopoldseder, S., Claren, J., Jurgens, C. & Sterner, R. (2004). Interconverting the catalytic activities of $(\beta\alpha)_8$-barrel enzymes from different metabolic pathways: sequence requirements and molecular analysis. J Mol Biol **337:** 871-879

Li, H., Robertson, A. D. & Jensen, J. H. (2005). Very fast empirical prediction and rationalization of protein pKa values. Proteins **61:** 704-721

Liang, S. & Grishin, N. V. (2004). Effective scoring function for protein sequence design. Proteins **54:** 271-281

Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. Nature **423:** 185-190

Looger, L. L. & Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol **307:** 429-445

Looger, L. L. & Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol **307:** 429-445

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J Phys Chem B **102:** 3586-3616

Mackerell, A. D., Jr. (2004). Empirical force fields for biological macromolecules: overview and issues. J Comput Chem **25:** 1584-1604

Mangalam, H. (2002). The Bio* toolkits--a brief overview. Brief Bioinform **3:** 296-302

Marshall, S. A., Vizcarra, C. L. & Mayo, S. L. (2005). One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. Protein Sci **14:** 1293-1304

Marti-Renom, M. A., Madhusudhan, M. S. & Sali, A. (2004). Alignment of protein sequences by their profiles. Protein Sci **13:** 1071-1087

Mayo, S. L., Olafson, B. D. & Goddard, W. A. (1990). DREIDING: A Generic Force Field for Molecular Simulations. J Phys Chem B **94:** 8897-8909

Meiler, J. & Baker, D. (2006). ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins **65:** 538-548

Merkl, R. & Waack, S. (2002). Bioinformatik Interaktiv. Wiley-VCH **Buch:**

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equations of state calculations by fast computing machines. J Chem Phys **21:** 1087-1092

Milch, B., Marthi,B., Russell,S. (2004). BLOG: Relational Modeling with Unknown Objects. CML 2004 Workshop on Statistical Relational Learning and Its Connections to Other Fields**:**

Moreland, J. L., Gramada, A., Buzko, O. V., Zhang, Q. & Bourne, P. E. (2005). The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. BMC Bioinformatics **6:** 21

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol **15:** 285-289

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol **247:** 536-540

Nagano, N., Orengo, C. A. & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol **321:** 741-765

Nagatani, R. A., Gonzalez, A., Shoichet, B. K., Brinen, L. S. & Babbitt, P. C. (2007). Stability for function trade-offs in the enolase superfamily "catalytic module". Biochemistry **46:** 6688-6695

Nedas, K. A., Pilgrim, B. (2005). Munkres-Kuhn (Hungarian) Algorithm Clean Version: 0.11. URL:wwwspatialmaineedu/~kostas**:**

Notredame, C., Abergel, C. (2003). Using Multiple Alignment Methods to Assess the Quality of Genomic Data Analysis, in Bioinformatics and Genomes: Current Perspectives. Horizon Scientific Press**:** 30-50

Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol **302:** 205-217

Olsson, M. H., Mavri, J. & Warshel, A. (2006). Transition state theory can be used in studies of enzyme catalysis: lessons from simulations of tunnelling and dynamical effects in lipoxygenase and other systems. Philos Trans R Soc Lond B Biol Sci **361:** 1417-1432

Pierce, A., Spriet, J., Desmet, J. & Mayo, S. L. (2000). Conformational splitting: A more powerful criterion for dead-end elimination. J Comput Chem**:**

Pierce, N. A. & Winfree, E. (2002). Protein design is NP-hard. Protein Eng **15:** 779-782

Pokala, N. & Handel, T. M. (2004). Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. Protein Sci **13:** 925-936

Pokala, N. & Handel, T. M. (2005). Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J Mol Biol **347:** 203-227

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol **193:** 775-791

Poole, A. M. & Ranganathan, R. (2006). Knowledge-based potentials in protein design. Curr Opin Struct Biol **16:** 508-513

Powers, N. & Jensen, J. H. (2006). Chemically accurate protein structures: validation of protein NMR structures by comparison of measured and predicted pKa values. J Biomol NMR **35:** 39-51

Pujadas, G., Ramirez, F. M., Valero, R. & Palau, J. (1996). Evolution of β-amylase: patterns of variation and conservation in subfamily sequences in relation to parsimony mechanisms. Proteins **25:** 456-472

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. J Mol Biol **7:** 95-99

Riede, P. (2006). Versuch der Etablierung einer neuen katalytischen Aktivität auf dem Proteingerüst der Indolglycerolphosphat Synthase aus *Sulfolobus solfataricus*. Universität Regensburg.

Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Eng **12:** 85-94

Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B. & Ranganathan, R. (2005). Natural-like function in artificial WW domains. Nature **437:** 579-583

Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol **234:** 779-815

Sali, A. & Overington, J. P. (1994). Derivation of rules for comparative protein modeling from a database of protein structure alignments. Protein Sci **3:** 1582-1596

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins **9:** 56-68

Schmidt, D. M., Mundorff, E. C., Dojka, M., Bermudez, E., Ness, J. E., Govindarajan, S., Babbitt, P. C., Minshull, J. & Gerlt, J. A. (2003). Evolutionary potential of $(\beta\alpha)_8$-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. Biochemistry **42:** 8387-8393

Schramm, V. L. (2005). Enzymatic transition states and transition state analogues. Curr Opin Struct Biol **15:** 604-613

Schuttelkopf, A. W. & van Aalten, D. M. (2004). PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallogr D Biol Crystallogr **60:** 1355-1363

Scott, E. E., He, Y. A., Wester, M. R., White, M. A., Chin, C. C., Halpert, J. R., Johnson, E. F. & Stout, C. D. (2003). An open conformation of mammalian cytochrome P450 2B4 at 1.6-A resolution. Proc Natl Acad Sci U S A **100:** 13196-13201

Seitz, T., Bocola, M., Claren, J. & Sterner, R. (2007). Stabilisation of a (betaalpha)8-barrel protein designed from identical half barrels. J Mol Biol **372:** 114-129

Shah, P. S., Hom, G. K. & Mayo, S. L. (2004). Preprocessing of rotamers for protein design calculations. J Comput Chem **25:** 1797-1800

Shah, P. S., Hom, G. K., Ross, S. A., Lassila, J. K., Crowhurst, K. A. & Mayo, S. L. (2007). Full-sequence computational design and solution structure of a thermostable protein variant. J Mol Biol **372:** 1-6

Shatsky, M., Nussinov, R. & Wolfson, H. J. (2004). A method for simultaneous alignment of multiple protein structures. Proteins **56:** 143-156

Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. (1995). A relationship between protein stability and protein function. Proc Natl Acad Sci U S A **92:** 452-456

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol **213:** 859-883

Sippl, M. J. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Comput Aided Mol Des **7:** 473-501

Sippl, M. J. (1995). Knowledge-based potentials for proteins. Curr Opin Struct Biol **5:** 229-235

Sippl, M. J. & Flockner, H. (1996). Threading thrills and threats. Structure **4:** 15-19

Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H. & Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. Nature **437:** 512-518

Sonnhammer, E. L., Eddy, S. R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins **28:** 405-420

Sousa, S. F., Fernandes, P. A. & Ramos, M. J. (2006). Protein-ligand docking: current status and future challenges. Proteins **65:** 15-26

Sterner, R. & Höcker, B. (2005). Catalytic versatility, stability, and evolution of the $(\beta\alpha)_8$-barrel enzyme fold. Chem Rev **105:** 4038-4055

Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. Fold Des **3:** 253-258

Summers, N. L. & Karplus, M. (1989). Construction of side-chains in homology modelling. Application to the C-terminal lobe of rhizopuspepsin. J Mol Biol **210:** 785-811

Tynan-Connolly, B. M. & Nielsen, J. E. (2006). pKD: re-designing protein pKa values. Nucleic Acids Res **34:** W48-51

Valdar, W. S. (2002). Scoring residue conservation. Proteins **48:** 227-241

Velec, H. F., Gohlke, H. & Klebe, G. (2005). DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. J Med Chem **48:** 6296-6303

Voet, D. & Voet, J. (2002). Biochemie, New York

Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. J Mol Biol **299:** 789-803

Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res **34:** 1692-1699

Wang, G. & Dunbrack, R. L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics **19:** 1589-1591

Wang, G. & Dunbrack, R. L., Jr. (2005). PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res **33:** W94-98

Wang, R., Lu, Y. & Wang, S. (2003). Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem **46:** 2287-2303

Warshel, A., Sharma, P. K., Kato, M. & Parson, W. W. (2006). Modeling electrostatic effects in proteins. Biochim Biophys Acta **1764:** 1647-1676

Wierenga, R. K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. FEBS Lett **492:** 193-198

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. & Richardson, D. C. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol **285:** 1711-1733

Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol **285:** 1735-1747

Wright, H., Noda-Garcia, L., Ochoa-Leyva, A., Hodgson, D. A., Fulop, V. & Barona-Gomez, F. (2007). The structure/function relationship of a dual-substrate $(\beta\alpha)_8$-isomerase. Biochem Biophys Res Commun**:**

Wüthrich, K. (1995). NMR - this other method for protein and nucleic acid structure determination. Biological Crystallography **51:**

Xiang, J., Jung, J. Y. & Sampson, N. S. (2004). Entropy effects on protein hinges: the reaction catalyzed by triosephosphate isomerase. Biochemistry **43:** 11436-11445

Yanofsky, C. (2001). Advancing our knowledge in biochemistry, genetics, and microbiology through studies on tryptophan metabolism. Annu Rev Biochem **70:** 1-37

Yanofsky, C. (2003). Using studies on tryptophan metabolism to answer basic biological questions. J Biol Chem **278:** 10859-10878

Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Rothlisberger, D. & Baker, D. (2006). New algorithms and an in silico benchmark for computational enzyme design. Protein Sci **15:** 2785-2794

Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res **33:** 2302-2309

Zollars, E. S., Marshall, S. A. & Mayo, S. L. (2006). Simple electrostatic model improves designed protein sequences. Protein Sci **15:** 2014-2018