1
Biologische Grundlagen

In den folgenden Kapiteln beschäftigen wir uns meist mit Algorithmen, die Eigenschaften von Makromolekülen bewerten oder vergleichen. Für das Verständnis der Methoden und Modellierungsansätze benötigen wir einige wenige biologische Grundkenntnisse, die in diesem Kapitel eingeführt werden. Zu den wichtigsten molekularbiologischen Objekten gehören DNA, RNA und Proteine. Dies sind Moleküle, die jeweils aus einer Abfolge kleinerer Bausteine aufgebaut sind. Deren lineare Anordnung kann in Form einer Zeichenkette (Sequenz) angegeben werden. Sequenzen betrachten wir im folgenden Kapitel genauer.

Die *DNA* ist der wichtigste Datenträger in der Molekularbiologie. Es wurden Hochdurchsatzmethoden entwickelt, mit denen die Zusammensetzung der DNA, d. h. deren Sequenz, mit geringem Aufwand und in kürzester Zeit ermittelt werden kann. Deswegen werden mittlerweile bevorzugt Genomsequenzen bestimmt, da aus diesen die Komposition der anderen Makromoleküle abgeleitet werden kann. Die *Proteine* sind die wichtigsten Grundbausteine aller biologischen Zellen. Sie geben den Zellen oft ihre Struktur und sind in Form von Enzymen wichtige Komponenten der meisten Stoffwechselvorgänge. Die biologische Bedeutung der *RNA* hat in den letzten Jahren durch neue biochemische Befunde extrem zugenommen. Es ist klar geworden, dass RNA-Moleküle in erheblichem Ausmaß an Regulationsaufgaben beteiligt sind, was lange unbekannt war.

Die *in vivo* Funktion von DNA, RNA und Proteinen kann nur anhand der dreidimensionalen Molekülstruktur komplett verstanden werden. Aufgrund ihrer Vielfalt nimmt im Folgenden die Darstellung von Proteinarchitekturen einen breiteren Raum ein. Nach der Beschreibung typischer 3D-Strukturen beschäftigen wir uns mit einigen Eigenschaften und Prozessen, die in bioinformatischen Algorithmen von Bedeutung sind. Das Kapitel schließt mit einer Definition wichtiger Fachbegriffe.

1.1
DNA

Im bioinformatischen Kontext stehen Sequenzen in der Regel für die Abfolge einer kleinen, definierten Menge von Einzelbausteinen. DNA-Sequenzen sind

Bioinformatik, 3., vollständig überarbeitete und erweiterte Auflage. Rainer Merkl.

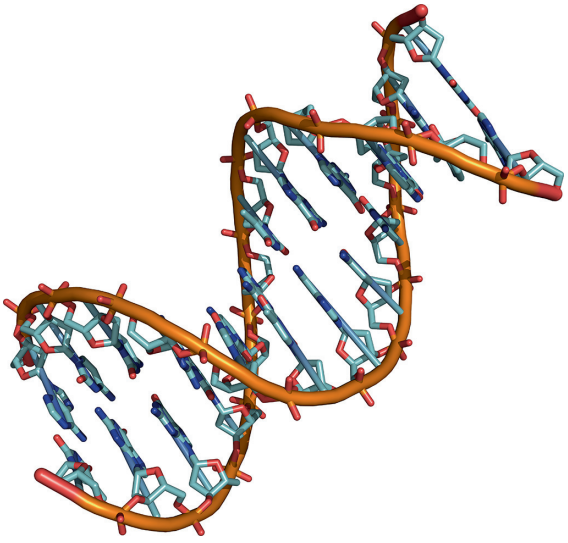

Abb. 1.1 Raumstruktur der DNA. In der Abbildung ist die Doppelhelix gut zu erkennen. Die basischen Anteile der Nukleotide sind nach innen gerichtet und durch Wasserstoffbrücken verknüpft. Außen verlaufen die Zucker-Phosphat-Anteile der polymerisierten Nukleotide.

Modelle für Makromoleküle der Desoxyribonukleinsäure (abgekürzt DNS oder DNA), die als fädige Struktur vorliegt. Jeder Strang ist eine Folge von vier Einzelbausteinen (Nukleotide), diese bestehen jeweils aus

- einem Zucker (in der DNA: Desoxyribose)
- einer der Purin- oder Pyrimidinbasen Adenin, Guanin oder Cytosin, Thymin
- einem Phosphatrest

In der Zelle kommt DNA üblicherweise in doppelsträngiger Form vor, die eine Doppelhelix bildet. In der Helix stehen sich Nukleotide paarweise gegenüber, wobei nur zwei Paarungen zugelassen sind (siehe Abb. 1.1 und 1.2).

Wasserstoffbrücken Die Funktion und Struktur von Makromolekülen wird maßgeblich durch *Wasserstoffbrücken* determiniert. Eine Wasserstoffbrücke ist eine anziehende elektromagnetische Wechselwirkung zwischen einem kovalent in einem Molekül gebundenen Wasserstoff und einem elektronegativen Atom wie Stickstoff oder Sauerstoff. Diese „Bindung" kann im Gegensatz zu einer kovalenten Atombindung mit relativ geringem Energieaufwand gelöst werden.

Reverses Komplement Aufgrund des chemischen Aufbaus der Nukleotide hat jeder DNA-Strang beliebiger Länge eine eindeutige Orientierung, mit jeweils einem freien 3′-OH- und einem 5′-OH-Ende. Sequenzen werden nach Übereinkunft stets so geschrieben, dass das 5′-OH Ende links und das 3′-OH-Ende rechts steht. *In vivo* ist die DNA-Doppelhelix meist zu einem Ring geschlossen, z. B. in Chromosomen oder Plasmiden. Darin sind die beiden komplementären DNA-

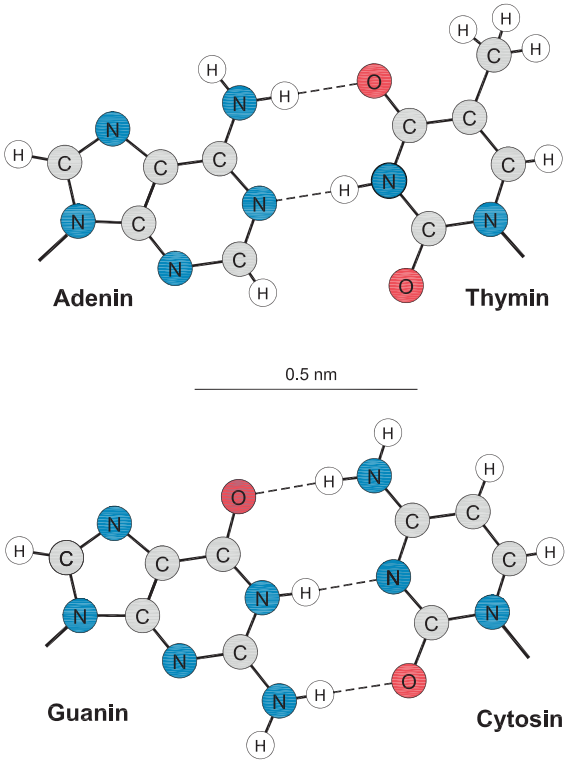

Abb. 1.2 Basenpaarungen in der DNA. In der als Doppelhelix bekannten DNA-Struktur liegen sich jeweils paarweise die Basen Adenin und Thymin beziehungsweise Guanin und Cytosin gegenüber. Zwischen AT-Paaren können zwei, zwischen GC-Paaren drei, Wasserstoffbrücken ausgebildet werden. Je höher der Anteil von GC-Paaren, desto mehr Energie muss für das Trennen der beiden Stränge einer DNA-Doppelhelix aufgewendet werden.

Stränge gegenläufig angeordnet. Die durch den Aufbau vorgegebene Orientierung bedingt die Richtung, in der Gene abgelesen werden. Da Gene auf beiden Strängen codiert sein können, in Datensammlungen jedoch nur die Sequenz eines Stranges abgelegt wird, muss zur Bestimmung der Sequenz des Gegenstranges das *reverse Komplement* gebildet werden. In den Zellkernen höherer Arten ist die DNA um Nukleosomen gewickelt, die sich zu komplexeren Strukturen zusammenlagern. Dieser Befund ist für die bioinformatischen Kernalgorithmen ohne Belang.

1.2 Genetischer Code und Genomkomposition

Die Sequenzinformation eines jeden Proteins ist in Form eines Gens in der DNA-Sequenz codiert. Jeweils drei direkt aufeinanderfolgende Nukleotide, die nicht überlappend abgelesen werden, codieren für eine Aminosäure. Eine solche Nu-

Tab. 1.1 Der genetische Code. Die Zahlen geben die Nukleotidposition im Codon an. In einigen speziellen Fällen, wie in mitochondrialen Genomen, kann es Abweichungen von diesem kanonischen Code geben. Die Namen der Aminosäuren sind im Dreibuchstabencode angegeben; siehe folgendes Kapitel.

		2				
		T	**C**	**A**	**G**	
1	**T**	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	**T**
		TTC Phe	TCC Ser	TAC Tyr	TGC Cys	**C**
		TTA Leu	TCA Ser	TAA Stop	TGA Stop	**A**
		TTG Leu	TCG Ser	TAG Stop	TGG Trp	**G**
	C	CTT Leu	CCT Pro	CAT His	CGT Arg	**T**
		CTC Leu	CCC Pro	CAC His	CGC Arg	**C**
		CTA Leu	CCA Pro	CAA Gln	CGA Arg	**A**
		CTG Leu	CCG Pro	CAG Gln	CGG Arg	**G**
	A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	**T**
		ATC Ile	ACC Thr	AAC Asn	AGC Ser	**C**
		ATA Ile	ACA Thr	AAA Lys	AGA Arg	**A**
		ATG Met	ACG Thr	AAG Lys	AGG Arg	**G**
	G	GTT Val	GCT Ala	GAT Asp	GGT Gly	**T**
		GTC Val	GCC Ala	GAC Asp	GGC Gly	**C**
		GTA Val	GCA Ala	GAA Glu	GGA Gly	**A**
		GTG Val	GCG Ala	GAG Glu	GGG Gly	**G**

3

kleotidgruppe wird Triplett oder *Codon* genannt. Die Abbildung der 64 Tripletts auf die 20 Aminosäuren heißt genetischer Code, dieser ist in Tab. 1.1 dargestellt. Der Code ist quasi universell, abweichende Codonzuordnungen finden sich aber z. B. bei Mitochondrien, *Mycoplasma* und einigen Protozoen (Übersicht in [1]).

Leseraster Die Struktur der DNA legt die Lage der einzelnen Gene innerhalb einer DNA-Sequenz nicht fest. Daher ergeben sich – wegen der zwei möglichen Ableserichtungen und der drei möglichen Intervalle pro Leserichtung – insgesamt sechs Leseraster. Prinzipiell kann jede Codonsequenz ein Gen codieren, sofern sie mit einem Startcodon beginnt und mit einem Stoppcodon endet. Eine derartige Sequenz wird zur Unterscheidung von Genen (für die eine Funktion nachgewiesen ist) offenes Leseraster (*open reading frame*, ORF) genannt.

Diese Situation wird im folgenden Beispiel klar; siehe Abb. 1.3. Je nach Leseraster resultieren aus derselben DNA-Sequenz unterschiedliche Proteinsequenzen. Im gezeigten Beispiel existiert genau ein ORF (hier im Leseraster 1), dessen Lage durch ein Startcodon (Met) und ein Stoppcodon (durch *** markiert) definiert ist; in allen anderen Leserastern treten in der gezeigten Sequenz Stoppcodons auf oder es fehlt ein Startcodon. Gene haben allerdings in der Regel eine Länge von mehr als 80 Codonen.

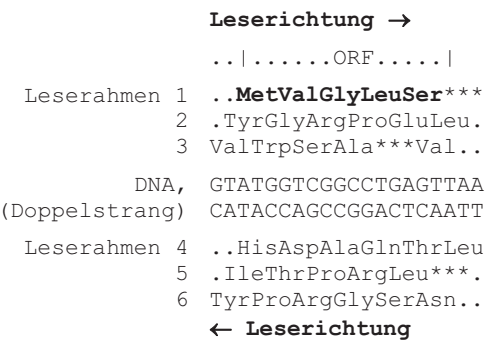

Abb. 1.3 Übersetzen eines DNA-Fragments in Proteinsequenzen. DNA kann in sechs Leserastern interpretiert werden. Ein ORF ist eine DNA-Teilsequenz, die durch ein Start- und ein Stoppcodon flankiert wird.

Informationsgehalt der Basenpositionen Der Informationsgehalt I der drei Basenpositionen im Codon ist nicht gleich, es gilt $I(\textit{Position 2}) > I(\textit{Position 1}) > I(\textit{Position 3})$ [2]. Hierfür ist der genetische Code verantwortlich: Eine Mutation der dritten Base im Codon verändert die Aminosäurenkomposition häufig nicht, eine Mutation in der ersten Basenposition führt häufig zum Einbau einer Aminosäure mit ähnlichen Eigenschaften, eine Mutation der mittleren Base verursacht häufig den Einbau einer Aminosäure mit anderen Eigenschaften [1]. Die geringsten Auswirkungen auf die Aminosäurenkomposition der Proteine haben somit Veränderungen der Basenkomposition in Position drei des Codons, gefolgt von Veränderungen der Basenkomposition an Position eins. Diese Befunde machen deutlich, dass simple statistische Konzepte nicht dazu geeignet sind, codierende Sequenzen adäquat zu modellieren. Es kann nicht unterstellt werden, dass die Basen voneinander unabhängig in Genen auftreten.

GC-Gehalt von Genomen Der GC-Gehalt, d. h. der relative Anteil von Guanin oder Cytosin an der DNA, ist eine charakteristische Größe eines Genoms. In bakteriellen Genomen schwankt der GC-Gehalt zwischen 25 und 75 %. In GC-Basenpaaren werden drei Basenpaarungen ausgebildet, in AT-Basenpaaren nur zwei; daher wurde lange vermutet, dass ein hoher GC-Gehalt des Genoms z. B. für thermophile [3] oder halophile [4] Organismen vorteilhaft wäre. Thermophile Organismen leben in Habitaten mit erhöhten Umgebungstemperaturen, halophile kommen in Umgebungen mit erhöhter Salzkonzentration vor. Es hat sich jedoch herausgestellt, dass der mittlere GC-Gehalt nicht von solchen Umweltfaktoren abhängt, sondern wohl durch evolutionären Druck eingestellt wird [5]. Zudem hängt der GC-Gehalt von Eigenschaften des DNA-Replikationssystems ab, dessen Aufgabe es ist, Kopien des Erbguts für die nächste Generation herzustellen. Aus dem Vergleich des GC-Gehalts der Genome solcher Bakteriophagen, die ihr eigenes DNA-Replikationssystem, und solcher, die das Replikationssystem des Wirts *Escherichia coli* verwenden, mit dem GC-Gehalt des Genoms von *Escheri-*

chia coli wurde geschlossen, dass der GC-Gehalt vom DNA-Replikationssystem moduliert wird [1]. Bestimmte Mutationen im *mutT* Gen von *Escherichia coli* induzieren Transversionen von AT- nach GC-Basenpaaren [6] und Mutationen im *mutY* Gen-Transversionen von GC- nach AT-Basenpaaren [7]. Die Genprodukte beider Gene sind an der DNA-Replikation bzw. DNA-Reparatur beteiligt. Interessanterweise gibt es aber definierte Bereiche in RNA-Molekülen, deren GC-Gehalt auf die optimale Wachstumstemperatur schließen lässt [8].

Codonhäufigkeiten Codonen kommen nicht mit annähernd gleicher Häufigkeit in Genen vor. Im Gegenteil, die Codonhäufigkeiten schwanken zwischen den taxonomischen Gruppen beträchtlich. Die Codonpräferenzen der beiden nahe verwandten Bakterien *Escherichia coli* und *Salmonella typhimurium* sind sich relativ ähnlich. Codonhäufigkeiten des Bakteriums *Bacillus subtilis*, das zu beiden eine große phylogenetische Distanz aufweist, sind auffällig anders. Solche Unterschiede können, wie wir später sehen werden, dazu genutzt werden, Gensequenzen unbekannter Herkunft einer biologischen Art zuzuweisen.

Synonyme Codonen Codonen, die für dieselbe Aminosäure codieren, werden *synonyme Codonen* genannt. Synonyme Codonen treten ebenfalls nicht mit vergleichbarer Häufigkeit auf, einige werden bevorzugt eingebaut. Daraus resultierende Unterschiede in der Häufigkeitsverteilung von kurzen Nukleotidketten können unter Verwendung statistischer Verfahren (Markov-Ketten) ausgenutzt werden, um die Lage von Genen vorherzusagen (z. B. im Programm Glimmer [9]). In Korrelation mit den ungleichmäßigen Codonhäufigkeiten treten Unterschiede in den speziesspezifischen tRNA-Konzentrationen auf. tRNA ist an der Translation, d. h. der RNA-instruierten Proteinsynthese, beteiligt.

Der genetische Code wird als *degeneriert* (im Sinne der in der Atomphysik eingeführten Bedeutung) bezeichnet, da einige Aminosäuren durch mehrere (synonyme) Codonen codiert werden.

Bevorzugte Codonen Bei manchen Spezies variieren Codonhäufigkeiten zudem stark zwischen einzelnen Genen [10]. In bestimmten Genen tritt speziesspezifisch eine Teilmenge der Codonen bevorzugt auf (Übersichten in [11, 12]). Diese Verzerrung der Codonhäufigkeiten (*codon usage bias*) ist positiv korreliert mit der Genexpression [13]. Mögliche Ursachen für diese Verzerrung der Codonhäufigkeiten sind die unterschiedlichen Konzentrationen der tRNAs [14, 15], das Aufrechterhalten der maximalen Elongationsrate, die Kosten für das Korrekturlesen sowie unterschiedliche Translationsraten der Codonen [16]. Diese Verzerrung der Codonhäufigkeiten wird als „Strategie“ interpretiert, die Wachstumsraten zu optimieren [11]. Wie wir später sehen werden, sind Unterschiede in den Codonhäufigkeiten ein wichtiges Signal, das für bioinformatische Analysen genutzt wird. Bei Prokaryonten weisen Gene, die im Genom benachbart liegen, eine ähnliche *codon usage* auf. Es wurde gezeigt, dass aus der Ähnlichkeit von Codonhäufigkeiten eine Interaktion der Genprodukte vorhergesagt werden kann [17]. Zudem belegen diese Befunde die komplexe Komposition codierender DNA-Sequenzen.

Tab. 1.2 Gemittelte Codonhäufigkeiten im Genom von *Escherichia coli* K-12. Die Summe der Prozentwerte ergibt 100.

		2				
		T	**C**	**A**	**G**	
1	**T**	TTT 2,08	TCT 0,89	TAT 1,53	TGT 0,49	**T**
		TTC 1,78	TCC 0,90	TAC 1,30	TGC 0,65	**C**
		TTA 1,22	TCA 0,64	TAA 0,19	TGA 0,09	**A**
		TTG 1,28	TCG 0,86	TAG 0,02	TGG 1,48	**G**
	C	CTT 1,00	CCT 0,65	CAT 1,23	CGT 2,29	**T**
		CTC 1,06	CCC 0,47	CAC 1,04	CGC 2,30	**C**
		CTA 0,35	CCA 0,81	CAA 1,43	CGA 0,32	**A**
		CTG 5,56	CCG 2,47	CAG 2,93	CGG 0,49	**G**
	A	ATT 2,91	ACT 0,91	AAT 1,58	AGT 0,76	**T**
		ATC 2,64	ACC 2,42	AAC 2,28	AGC 1,59	**C**
		ATA 0,36	ACA 0,59	AAA 3,47	AGA 0,16	**A**
		ATG 2,80	ACG 1,37	AAG 1,07	AGG 0,11	**G**
	G	GTT 1,88	GCT 1,57	GAT 3,18	GGT 2,60	**T**
		GTC 1,49	GCC 2,51	GAC 2,05	GGC 3,07	**C**
		GTA 1,11	GCA 1,98	GAA 4,12	GGA 0,67	**A**
		GTG 2,66	GCG 3,49	GAG 1,80	GGG 1,02	**G**

3

Codon usage von Escherichia coli K-12 In Tab. 1.2 sind die gemittelten Codonhäufigkeiten angegeben, so wie sie im Genom des Bakteriums *Escherichia coli* K-12 vorkommen. Auffallend selten sind in diesem Genom die Codonen AGA, AGG und CTA.

1.3 Transkription

Die unmittelbar verwendete Datenbasis für die biologische Proteinsynthese ist nicht die Sequenz der DNA, sondern die eines *messenger* RNA (mRNA) Moleküls, das als Kopie eines Genabschnittes hergestellt wird. Ganz allgemein wird das Umschreiben eines Textes *Transkription* genannt. In Analogie hierzu wird die Produktion dieser mRNA ebenso bezeichnet. Die für die Transkription notwendigen Enzyme sind die DNA-abhängigen RNA-Polymerasen. Bei der Transkription wird, anstelle von T (Thymin), in die mRNA das Nukleotid U (Uracil) eingebaut. Das RNA-Molekül, das hierbei entsteht, wird *Transkript* genannt.

Bei der RNA-Synthese müssen zwei Bedingungen eingehalten werden:

- Die Synthese muss unmittelbar vor einem Gen beginnen.
- Es muss der sinntragende (codogene) Strang transkribiert werden.

Das Einhalten dieser Bedingungen wird erreicht durch die bevorzugte Bindung von RNA-Polymerase an Erkennungsstellen (*Promotoren*), die unmittelbar vor

Abb. 1.4 Konsensussequenz von *Escherichia coli* Promotoren. Der untere der beiden DNA-Stränge wird transkribiert ab Position +1; nach [18].

Genen liegen. Bei der Transkription lagern sich an den codogenen Strang komplementäre Ribonukleotide an, sodass z. B. aus der Sequenz TAC das Startcodon AUG wird.

Promotoren am Beginn des Transkriptes Vergleicht man die Promotoren von *Escherichia coli* und bildet hieraus einen „idealen Promotor", so fällt Folgendes auf:

- In einem Bereich, der circa zehn Basenpaare stromaufwärts des Transkriptionsstarts liegt, findet sich eine Sequenz, die häufig ähnlich zu TATA (-10-Region oder *TATA-Box*) ist.
- In einem Bereich, der circa 35 Basenpaare stromaufwärts vom Start liegt (-35-Region), befindet sich innerhalb eines AT-reichen Abschnittes eine Sequenz, die häufig ähnlich zu TTGACA ist.

Abbildung 1.4 zeigt einen idealisierten Promotor; von dessen Zusammensetzung weichen bekannte Promotoren mehr oder weniger stark ab.

Funktion von Transkriptionsfaktoren Für die Einleitung der Transkription ist es notwendig, dass Transkriptionsfaktoren an den Promotor oder an zusätzliche Bindestellen wie *Enhancer* binden. In vielen Fällen ist das genaue Zusammenwirken dieser Faktoren nicht bekannt. Das Erkennen von Promotoren und anderen Bindestellen in DNA-Sequenzen ist eine wichtige Aufgabe der Bioinformatik.

Die Funktion des Operons In prokaryontischen Genomen sind Gene häufig in Funktionseinheiten, den *Operons*, zusammengefasst. Diese bestehen aus einem Promotor und einer Menge von Genen. Deren Genprodukte sind meist Elemente einer größeren Funktionseinheit oder tragen zur selben Stoffwechselleistung bei. So finden sich die Gene, die an der Tryptophanbiosynthese beteiligt sind, in einem Operon. Das Identifizieren von Promotoren mittels bioinformatischer Methoden hilft, Operons mit höherer Sicherheit vorherzusagen.

1.4 RNA

Bei höheren Eukaryonten kennt man nur für einen kleinen Bruchteil des Genoms die genaue Funktion [19]. Zu den Genomabschnitten mit bekannter Funktion gehören regulatorische Elemente wie Promotoren sowie die Gene, die für Proteine

oder bestimmte RNA-Spezies codieren. Für die RNA war lange Zeit eine Funktion als Transfer-RNA, als Komponente von Ribosomen (ribosomale RNA) oder von Spleißosomen gesichert. Der erheblich größere Rest des Genoms wurde häufig als *Junk DNA* bezeichnet. Jüngste, genomweite Experimente im Rahmen des ENCODE-Projektes haben jedoch gezeigt, dass Tausende, nicht für Proteine codierende, Transkripte (ncRNAs) existieren, deren Bedeutung unklar ist. Diese Ergebnisse belegen für das Genom des Menschen [20] und der Maus, dass der größte Teil transkribiert wird. ncRNAs werden in kleine interferierende RNAs, mikroRNAs und lange ncRNAs eingeteilt. Letztere haben eine Länge von mehr als 200 Nukleotiden und stellen den größten Anteil. Für diese RNA-Moleküle ist eine Beteiligung an der Organisation der Genomarchitektur und der Genexpression plausibel. Kleine RNA-Moleküle sind an einer Vielzahl von posttranskriptionalen *silencing*-Mechanismen beteiligt. Diese Prozesse zerstören mRNA-Moleküle, sodass kein Genprodukt (in der Regel ein Protein) gebildet werden kann.

1.5 Proteine

Proteine sind ebenfalls lineare Makromoleküle; Sonderfälle, die vom linearen Aufbau abweichen, sind für uns nicht von Belang. Bausteine sind in diesem Fall die 20 natürlich vorkommenden Aminosäuren. Der Aufbau dieser Molekülfamilie ist einheitlich und besteht aus einem, in allen Aminosäuren identischen, sowie einem variablen Teil, der häufig auch *Aminosäurerest* oder *Residuum* genannt wird (siehe Abb. 1.5). Form und Art dieses Restes beeinflussen die Wechselwirkungen zwischen den Bausteinen. Die wichtigsten Wechselwirkungen sind Wasserstoffbrückenbindungen zwischen polaren Seitenketten.

Natur der Aminosäuren Aufgrund des unterschiedlichen Aufbaus der Seitenkette haben die Aminosäuren voneinander abweichende physikalisch-chemische Eigenschaften. Sie lassen sich z. B. bezüglich der ionischen Ladung in die Gruppen *basisch*, *sauer* und *neutral* einteilen. Unter den neutralen Aminosäuren, die keine elektrische Gesamtladung tragen, finden sich wiederum *polare*, d. h. solche, die innerhalb des Moleküls eine unterschiedliche Ladungsverteilung aufweisen.

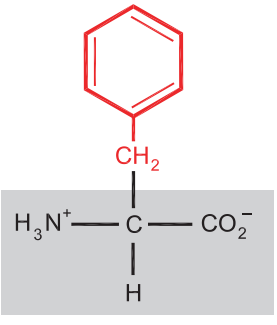

Abb. 1.5 Strukturformel der Aminosäure Phenylalanin. Der in allen Aminosäuren gleichartige Anteil ist in der Strukturformel grau unterlegt. In jeder Aminosäure ist mit dem zentralen C-Atom ein Wasserstoffatom (unten), eine Aminogruppe (links), eine Carboxylgruppe (rechts) und eine Seitengruppe (oben) verknüpft. Das zentrale C-Atom wird wegen seiner Lage im Molekül häufig als C_α-Atom bezeichnet.

Tab. 1.3 Vorkommen der Aminosäuren in Proteinen. Die Werte sind in Prozent angegeben und wurden aus einer repräsentativen Stichprobe ermittelt; nach [21]. Der hier verwendete Einbuchstabencode lautet wie folgt: A, Alanin; C, Cystein; D, Asparaginsäure; E, Glutaminsäure; F, Phenylalanin; G, Glycin; H, Histidin; I, Isoleucin; K, Lysin; L, Leucin; M, Methionin; N, Asparagin; P, Prolin; Q, Glutamin; R, Arginin; S, Serin; T, Threonin; V, Valin; W, Tryptophan; Y, Tyrosin.

Aminosäure	**A**	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
Häufigkeit [%]	8,66	4,40	3,91	5,70	1,93	3,67	5,81	8,33	2,44	4,85	8,62	6,20	1,95	3,84	4,58	6,95	6,10	1,44	3,53	7,09

Apolare, neutrale Aminosäuren sind *hydrophob* (wasserabstoßend). Sie tendieren dazu, untereinander und mit anderen hydrophoben Gruppen zu interagieren. Mit *hydrophil* werden Moleküle und Residuen bezeichnet, die gut wasserlöslich sind. Ein Spezialfall ist Prolin, eine zyklische Aminosäure. Nach der Ausbildung der Peptidbindung steht in dieser Aminosäure kein Wasserstoff mehr zur Ausbildung von Wasserstoffbrückenbindungen zur Verfügung. Diese Eigenart hat erheblichen Einfluss auf die Proteinstruktur.

Die Häufigkeiten, mit denen die 20 Aminosäuren in Proteinen vorkommen, unterscheiden sich deutlich. In Tab. 1.3 ist das mittlere Vorkommen gelistet.

Die in Abb. 1.6 dargestellten Verwandtschaftsbeziehungen aufgrund physikalischer und chemischer Eigenschaften der Aminosäuren sind die Grundlage für viele Sequenzvergleichs- und Alignmentverfahren. Hierfür werden Scoring-Matrizen benötigt, die wiederum aus Substitutionshäufigkeiten bestimmt werden. Diese Häufigkeiten werden aus dem Vergleich einer Vielzahl ähnlicher Proteine ermittelt und spiegeln gemeinsame Eigenschaften von Aminosäuren wider. Die angesprochenen Verfahren und Datensätze werden in den folgenden Kapiteln genauer vorgestellt.

1.6 Peptidbindung

Proteine sind Polypeptidketten, die aus Aminosäuren synthetisiert werden. Bei der Synthese wird die Carboxylgruppe (COOH) der einen Aminosäure mit der Aminogruppe (NH_2) des Nachbarn durch eine kovalente Bindung (Peptidbindung) verknüpft. Jede Polypeptidkette beliebiger Länge hat ein freies Amino-Ende (N-Terminus) und ein freies Carboxyl-Ende (C-Terminus). Die Richtung einer Kette ist definiert als vom N-Terminus zum C-Terminus zeigend. Diese Richtung stimmt überein mit der Syntheserichtung *in vivo*, die mit dem Ablesen der mRNA in 5′-3′-Richtung korrespondiert.

ϕ- und ψ-Winkel Die an der Peptidbindung beteiligten Atome liegen jeweils starr in einer Ebene. Daher wird der *Hauptkettenverlauf* einer Polypeptidkette durch die Angabe von zwei Winkeln (ϕ, ψ) pro Residuum beschrieben. Diese Winkel

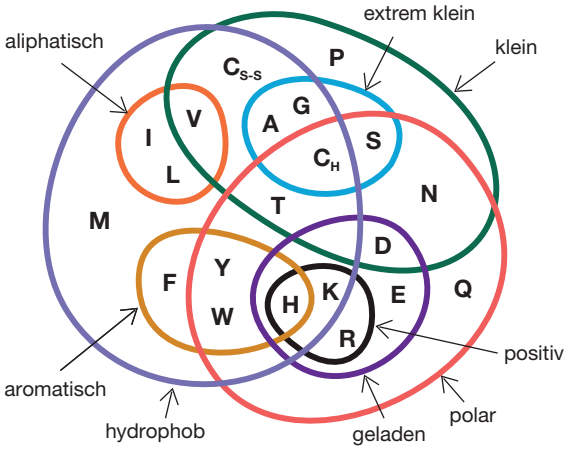

Abb. 1.6 Venn-Diagramm der 20 natürlichen, in Proteinen vorkommenden Aminosäuren. Die Aminosäuren wurden aufgrund solcher physikalisch-chemischer Eigenschaften gruppiert, die für die Tertiärstruktur von Proteinen wichtig sind. Die Aminosäuren sind im Wesentlichen in zwei Gruppen (polar und hydrophob) eingeteilt, eine dritte Gruppe (klein) umfasst die kleinen Aminosäuren. Die Menge extrem klein enthält diejenigen Aminosäuren, die höchstens zwei Seitenkettenatome besitzen. Cystein (C) in reduzierter Form (C_H) ist Serin (S) ähnlich, in oxidierter Form (C_{S-S}) ähnelt es Valin (V). Aufgrund des speziellen Einflusses auf den Hauptkettenverlauf liegt Prolin (P) isoliert; nach [22]. Der Einbuchstabencode wird im folgenden Kapitel genauer erläutert und ist in der Legende zu Tab. 1.3 angegeben.

geben die Drehung der beiden, am Hauptkettenverlauf beteiligten Bindungen des zentralen C_α-Atoms jeder Aminosäure an. Beide Winkel unterliegen weiteren Einschränkungen, die sich aus der Natur des jeweiligen Aminosäurerestes herleiten. Die Rigidität der Peptidbindung und die sterische Hinderung zwischen Haupt- und Seitenkette tragen zur Stabilisierung der Proteinkonformation bei. Das erste Kohlenstoffatom, das im Rest auf das C_α-Atom folgt, wird C_β-Atom genannt. In Abb. 1.7 ist die Situation illustriert. Der Hauptkettenverlauf dient häufig dazu, Faltungstypen von Proteinen zu charakterisieren und zu vergleichen. Die Hauptkette heißt im Englischen *backbone*.

1.7 Konformation von Aminosäureseitenketten

Die Aminosäuren unterscheiden sich in der Art ihrer Seitenketten. Diese sind unterschiedlich lang und von verschiedener chemischer Natur. Jede Seitenkette kann eine von mehreren *Konformationen* einnehmen, die auf die Rotationsmöglichkeiten der Atombindungen zurückzuführen sind. Jede Konformation wird durch die Rotationswinkel beschrieben, die an den drehbaren Bindungen auftreten. Für die Zwecke des Proteindesigns, d. h. die rechnergestützte Modellierung, wird aus Komplexitätsgründen eine beschränkte Menge aller möglicher Seiten-

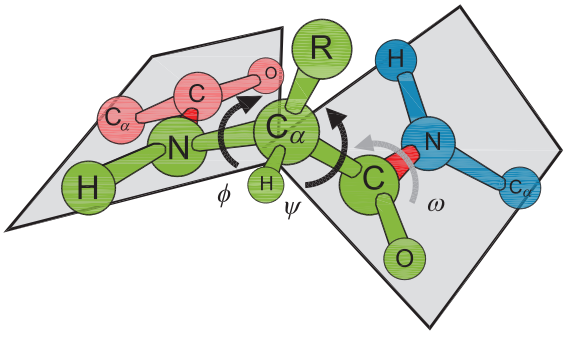

Abb. 1.7 Konformation der Peptidbindung. Die an einer Peptidbindung beteiligten sechs Atome liegen jeweils in einer Ebene. In der Abbildung sind zwei derartige Bindungen gezeigt und rot markiert. Der Aminosäurerest an der betrachteten Position (hier grün) ist mit R bezeichnet. Die räumliche Anordnung des Hauptkettenverlaufes eines Polypeptids …$-C_\alpha-C-N-C_\alpha-C-N-C_\alpha-$…wird bestimmt durch das, für jede Position (jedes Residuum) anzugebende, Paar von Winkeln (ϕ, ψ). Mit diesem Paar ist die Lage der durch die Peptidbindung aufgespannten Flächen relativ zum C_α-Atom festgelegt. Der mit ω bezeichnete Winkel kann nur die Werte +180° oder −180° annehmen.

kettenkonformationen betrachtet, die *Rotamere* genannt werden. Diese sind in Bibliotheken zusammengefasst [23, 24] und enthalten diejenigen Konformationen, die in Proteinen häufig vorkommen. Aufgrund der unterschiedlichen Anzahl rotierbarer Atombindungen ist die Dimension des Konformationsraumes abhängig von der betrachteten Aminosäure: Da die Seitenketten von Glycin und Alanin keine rotierbaren Bindungen aufweisen, genügt es, diese beiden Aminosäuren jeweils durch ein Rotamer zu repräsentieren. Die Seitenketten von Arginin und Lysin sind hingegen lang gestreckt. Mit vier rotierbaren Bindungen und drei energetisch günstigen Winkeln pro Bindung resultieren jeweils 81 Rotamere. Beispiele für Rotamere sind in Abb. 1.8 zusammengefasst. Die Menge der heute bekannten Proteinstrukturen erlaubt es, die Rotamerverteilungen in Abhängigkeit von den ϕ- und ψ-Winkeln der Hauptkette zu bestimmen. Solch hauptkettenspezifischen (*backbone dependent*) Bibliotheken [23, 25], verbessern die Modellierungsleistung beim Proteindesign.

1.8 Ramachandran-Plot

In Polypeptidketten sind nicht alle möglichen Kombinationen von ϕ- und ψ-Winkeln gleichhäufig. Wird die Verteilung dieser Winkel aus einer größeren Anzahl von Proteinen ermittelt, so ergeben sich die in der Abb. 1.9 gezeigten Präferenzen. Dieser Befund macht klar, dass im Konformationsraum nur drei Bereiche stärker besetzt sind. In idealisierter Weise fallen Residuen aus rechtsgängigen α-Helices in den Bereich von (−57°, −47°), während solche aus linksgängigen Helices bei (+57°, +47°) liegen. Residuen aus parallelen β-Faltblättern haben (ϕ,

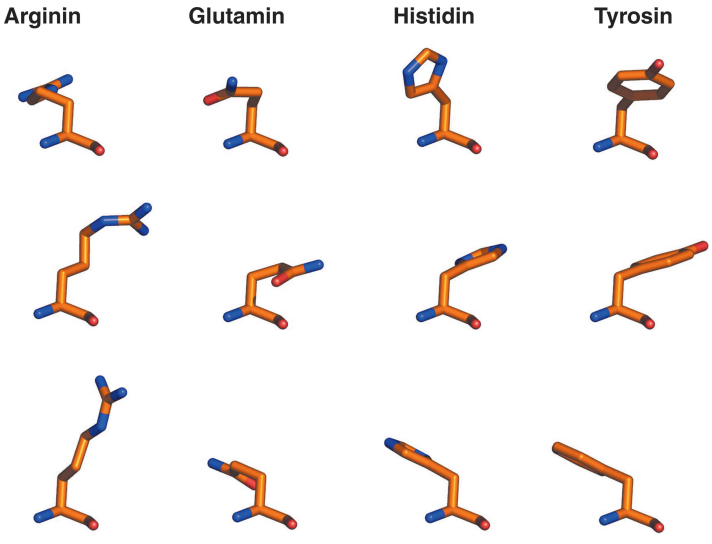

Abb. 1.8 Beispiele für Rotamerausprägungen. Rotamere sind in Proteinen häufig vorkommende Seitenkettenkonformationen. In der Abbildung sind für die Aminosäuren Arginin, Glutamin, Histidin und Tyrosin jeweils drei Rotamere angegeben. Die Seitenkette von Arginin enthält vier drehbare Bindungen mit jeweils drei energetisch günstigen Winkeln. Daher ergeben sich für Arginin 81 Rotamere (3^4). Für die Seitenkette von Glutamin resultieren aus drei drehbaren Bindungen 27 Rotamere. In den Seitenketten von Tyrosin und Histidin kommen jeweils nur zwei drehbare Bindungen vor, sodass neun Rotamere zur Beschreibung des Konformationsraumes ausreichen.

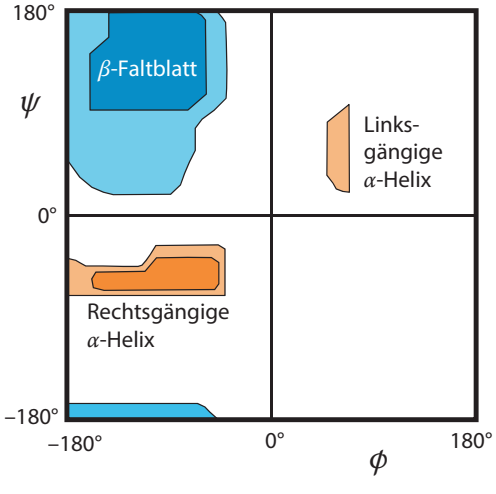

Abb. 1.9 Ramachandran-Plot. Je nach Zugehörigkeit zu einem Sekundärstrukturelement ergeben sich für die ϕ- und ψ-Winkel der Residuen charakteristische Kombinationen.

ψ)-Winkelkombinationen von circa (−119°, −113°), während diejenigen aus antiparallelen Blättern bei (−139°, +135°) zu finden sind. Werden für sämtliche Residuen eines Proteins die (ϕ, ψ)-Winkel bestimmt, so liegen häufig einige Paare abseits der Maxima. Dazu gehören solche von Glycin-Resten. Der Einbau von Glycin bewirkt eine scharfe Wendung des Hauptkettenverlaufs. Diese Darstellung der Winkelkombinationen wird nach ihrem Entwickler *Ramachandran-Plot* genannt. Die erwähnten Sekundärstrukturelemente werden im folgenden Text genauer erläutert.

1.9 Hierarchische Beschreibung von Proteinstrukturen

Die Eigenschaften der Seitenketten bestimmen die Wechselwirkungen innerhalb des Proteins und damit dessen dreidimensionale Konformation. K.U. Linderstrøm-Lang schlug 1952 vier Abstraktionsebenen vor, mit denen Proteine beschrieben werden können [26]. Dies sind:

- Die *Primärstruktur*, gebildet durch die Abfolge (Sequenz) der Aminosäuren.
- Die *Sekundärstruktur*: Aus der Polypeptidkette falten sich Sekundärstrukturelemente, die regelmäßige Arrangements des Hauptkettenverlaufes ergeben.
- Die *Tertiärstruktur*: Sie beschreibt die räumliche Anordnung aller Atome im Raum.
- Die *Quartärstruktur*: Sie definiert die Anordnung von Proteinen in Proteinkomplexen.

Wir werden Algorithmen kennenlernen, die darauf abzielen, Primär- Sekundär- und Tertiärstruktur von Proteinen zu analysieren, zu vergleichen oder vorherzusagen.

1.10 Sekundärstrukturelemente

Die Grundbausteine der Proteine sind die Aminosäuren. Deren Abfolge in Proteinen definiert die Proteinsequenz, d. h. die Primärstruktur. Die nächsthöhere Abstraktionsebene, auf der Proteine beschrieben werden können, ist die der *Sekundärstruktur*. Sekundärstrukturelemente sind regelmäßige 3D-Substrukturen des Hauptkettenverlaufes einer Peptidkette. Bei der Klassifizierung von Sekundärstrukturelementen werden Art und Anordnung der Aminosäurereste (Seitenketten) ignoriert. Die Stabilisierung der Sekundärstruktur erfolgt über Wasserstoffbrückenbindungen zwischen den Imino- und Carbonylgruppen *innerhalb der Hauptkette*.

Zusätzlich zu den hier beschriebenen Bindungskräften wird die 3D-Struktur eines Proteins im Wesentlichen durch schwache, nicht kovalente Wechselwirkungen der Aminosäureseitenketten, insbesondere durch Wasserstoffbrücken-

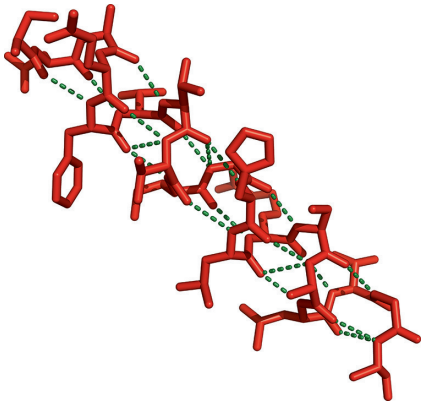

Abb. 1.10 Typische α-Helix. Wasserstoffbrücken sind gestrichelt eingezeichnet. Sie werden zwischen Atomen des Proteinrückgrates ausgebildet. Die Struktur ist hier als Stäbchenmodell gezeigt.

bindungen zwischen polaren Resten bestimmt. Diese Wechselwirkungen spielen bei der Betrachtung der Sekundärstruktur *keine* Rolle. Die beiden wichtigsten Sekundärstrukturelemente sind die α-Helix und das β-Faltblatt.

1.11 α-Helix

Sind die (ϕ, ψ)-Winkel aufeinander folgender Residuen konstant, so ergeben sich helikale Strukturen. Unter diesen ist die am häufigsten vorkommende die α-Helix. In der α-Helix besteht jeweils eine Wasserstoffbrückenbindung zwischen der CO-Gruppe einer Aminosäure und der NH-Gruppe der viertnächsten. Es machen jeweils 3,6 Aminosäuren eine vollständige Drehung aus. Die Abb. 1.10 zeigt einen typischen Vertreter einer α-Helix.

1.12 β-Faltblätter

Das zweite, wichtige Sekundärstrukturelement ist das β-Faltblatt. Ein β-Faltblatt besteht aus einzelnen β-Strängen, die meist 5–10 Residuen lang sind (siehe Abb. 1.11). In β-Faltblättern bilden sich Wasserstoffbrückenbindungen zwischen Residuen *unterschiedlicher* Stränge aus. Hierbei wechselwirken die C=O-Gruppen des einen Stranges mit den NH-Gruppen des nächsten Stranges. Auf diese Weise können mehrere Stränge ein Blatt bilden. Die C_α-Atome aufeinanderfolgender Residuen kommen abwechselnd über oder unter der Ebene, die durch das Faltblatt aufgespannt wird, zum Liegen. Die Stränge können in zwei Richtungen verlaufen:

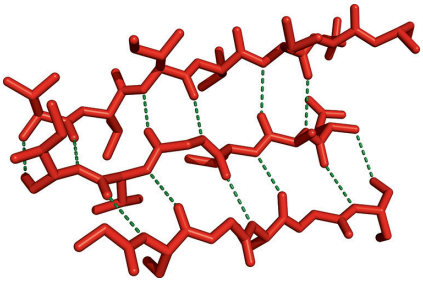

Abb. 1.11 β-Faltblatt bestehend aus drei Strängen. Wasserstoffbrücken sind gestrichelt eingezeichnet. Die Struktur ist als Stäbchenmodell dargestellt.

- *Parallel*; die durch N- und C-Terminus vorgegebene Richtung in nebeneinanderliegenden Strängen ist dieselbe.
- *Antiparallel*; die Richtung nebeneinanderliegender β-Stränge wechselt alternierend.

Im Proteininneren sind die β-Faltblätter meist parallel. An der Proteinoberfläche sind sie häufig antiparallel. Dort ragen die Aminosäurereste der einen Seite in die (hydrophile) Umgebung, während die der anderen zum hydrophoben Kern hin ausgerichtet sind. Hieraus ergibt sich im Idealfall in der Sequenz ein charakteristischer Wechsel von hydrophilen und hydrophoben Aminosäuren.

1.13 Supersekundärstrukturelemente

Die regulären Strukturen der Hauptkette werden ausgebildet, weil sie energetisch günstig sind. Sie bilden häufig Aggregate, die als Supersekundärstrukturelemente bezeichnet werden. So besteht der klassische Faltungstyp des $(\beta\alpha)_8$-Fasses beispielsweise aus acht $(\beta\alpha)$-Einheiten, die rotationssymmetrisch zur Mittelachse angeordnet sind. Die acht β-Stränge bilden eine fassartige Struktur, die außen von den α-Helices bedeckt wird. Das in Abb. 1.12 gezeigte Enzym HisF ist an der Histidinbiosynthese beteiligt. In HisF sind die acht $(\beta\alpha)$-Einheiten durch weitere Sekundärstrukturelemente ergänzt. Die Topologie des $(\beta\alpha)_8$-Fasses kommt in vielen Enzymfamilien vor, die völlig unterschiedliche Reaktionen katalysieren. Aus dieser breiten Verteilung auf völlig verschiedene Stoffwechselwege wurde gefolgert, dass dieser Faltungstyp bereits sehr früh in der Proteinevolution entstand [27]. Das auf der Erde vermutlich mengenmäßig häufigste Protein ist das Enzym Rubisco. Es ist an der Fotosynthese beteiligt und besitzt ebenfalls diese Topologie [28]. Ausführlich wird diese Faltungstopologie in [29, 30] beschrieben.

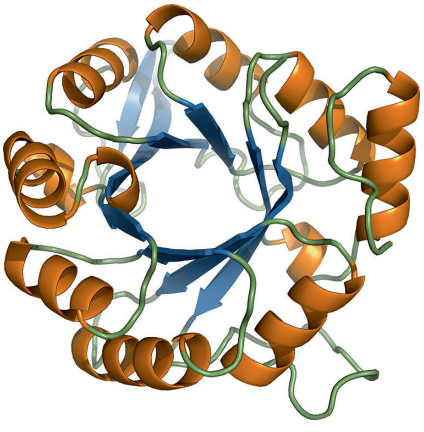

Abb. 1.12 Das $(\beta\alpha)_8$-Fass-Protein HisF. Beim Faltungstyp der $(\beta\alpha)_8$-Fässer bilden acht β-Stränge ein zentrales, in sich geschlossenes Faltblatt, das von acht α-Helices umgeben ist. Diese idealisierte Struktur ist häufig durch zusätzlich Schleifen oder andere Sekundärstrukturelemente erweitert.

1.14 Proteindomänen

Beim Vergleich zweier verwandter Proteinsequenzen fällt häufig auf, dass die Sequenzähnlichkeit nicht über die gesamte Länge hinweg einen konstant hohen Wert aufweist. Häufig wechseln sich Regionen mit signifikant hohen Scores (einem Maß für Sequenzähnlichkeit) ab mit solchen Regionen, die keinerlei Ähnlichkeit zur Vergleichssequenz haben. Ursache für dieses Schwanken des Scores ist der modulare Aufbau von Proteinen aus Domänen.

Eine *Domäne* ist bei Proteinen die kleinste Einheit mit einer definierten und unabhängig gefalteten Struktur. Proteindomänen bestehen meist aus 50–150 Aminosäuren und führen häufig individuelle Reaktionen aus, deren Zusammenwirken die Gesamtfunktion eines Proteins ausmacht.

In Abb. 1.13 ist die 3D-Struktur eines CAP-Monomers dargestellt. Dieses besteht aus zwei Domänen:

- Die N-terminale Domäne (Residuen 1–135) bindet cAMP und ist an der Dimerisierung beteiligt.
- Die C-terminale Domäne (Residuen 136–209) vermittelt die DNA-Bindung des Proteins.

CAP-Dimere, d. h. Aggregate von zwei Monomeren, aktivieren in Bakterien Gene, deren Genprodukte in den Zuckerstoffwechsel eingreifen.

Domänen sind die Organisationseinheiten, deren Zusammenwirken die Funktion eines Proteins bestimmt. Einen Eindruck von der Variabilität der Proteine auf Domänenniveau vermittelt Abb. 1.14. Auf Domänenebene lassen sich die beiden Proteine SAP97 und MAGI-1A wie folgt beschreiben: Beide Prote-

Abb. 1.13 3D-Struktur eines CAP-Monomers. Die N-terminale Domäne wurde orange, die C-terminale Domäne wurde blau eingefärbt. *In vivo* lagern sich jeweils zwei CAP-Moleküle zu einem Dimer zusammen; nach [31].

ine enthalten eine GuKc-Domäne und eine unterschiedliche Anzahl von PDZ-Domänen. Die GuKc-Domäne besitzt in aktiven Enzymen Guanylatkinaseaktivität, in membranassoziierten Proteinen zeigt sie nur Proteinbindungsfunktion. Die PDZ-Domänen haben unterschiedliche Bindungsspezifitäten; manche binden C-terminale, andere interne Polypeptide. In MAGI-1A kommt zusätzlich die ww-Domäne zweimal, in SAP97 die SH3-Domäne einmal vor.

1.15 Proteinfamilien

Aus dem letzten Absatz könnte gefolgert werden, dass Proteine eine schier unendliche Diversität von Strukturen hervorgebracht haben. Dies ist jedoch nicht der Fall. Wir konzentrieren uns im Folgenden auf Domänen, die in Multidomänenproteinen kombiniert werden oder in Eindomänenproteinen den Faltungstyp spezifizieren. Eindomänenproteine stellen den größten Anteil der bekannten Proteine. Es wurde abgeschätzt, dass circa 80 % aller Proteine zu einem von circa 400 Faltungstypen gehören. Diese Faltungstypen werden jeweils durch eine Supersekundärstruktur charakterisiert. Proteine können aufgrund dieser Faltungstypen gruppiert werden. Im Kapitel zu Datenbanken wird das Klassifikationssystem SCOP [32] vorgestellt, das auf einem solchen Schema beruht. Wie sehen repräsentative Vertreter der Faltungstypen aus? In den Abb. 1.15–1.20 werden Beispiele für die wichtigsten Faltungstypen im *Cartoon*-Modus präsentiert, hierbei wird auf die Wiedergabe der Seitenketten verzichtet. Diese Darstellung

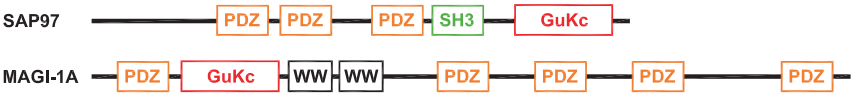

Abb. 1.14 Domänenstruktur des präsynaptischen Proteins SAP97 und des MAGI-1A Proteins.

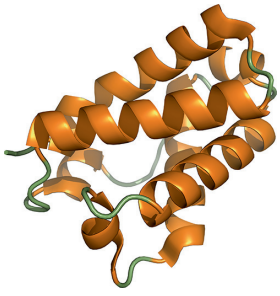

Abb. 1.15 Beispiel für ein *all-alpha*-Protein. Dieses Protein (1DLW) besitzt einen Globin-ähnlichen Faltungstyp. Die SCOP-Klassifikation lautet: sechs Helices, gefaltetes Blatt, teilweise geöffnet. In Klammern ist der Bezeichner angegeben, mit dem der Datensatz in der Strukturdaten-Bank PDB zu finden ist.

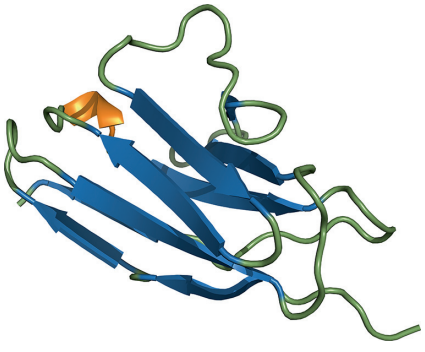

Abb. 1.16 Das Bence-Jones-Protein (1BWW) ist ein *all-beta*-Protein. Die SCOP-Klassifikation lautet: Sandwich, sieben Stränge in zwei Faltblättern, einige Mitglieder dieses Typs besitzen zusätzliche Stränge.

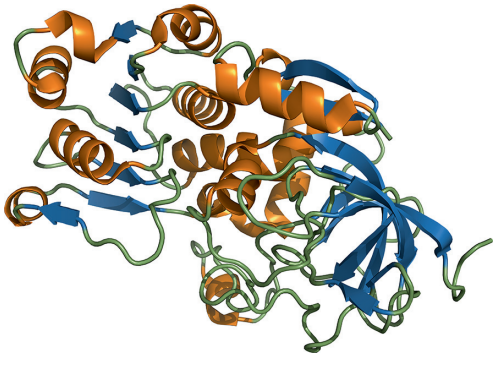

Abb. 1.17 Die NAD(P)-bindende Domäne des *Rossmann-folds* (2JHF) gehört zu den *alpha and beta folds* (*a*/*b*). Der Kern besteht aus drei Schichten, dazu kommt ein paralleles β-Faltblatt bestehend aus sechs β-Strängen.

des Rückgrates vermittelt die relative Anordnung der Sekundärstrukturelemente α-Helix, β-Strang und Schleife (*loop*).

Für die Klassifikation sind nur die α-Helix und der β-Strang von Belang. Aufgrund der Beschränkung auf zwei Klassifikationselemente existieren auch nur

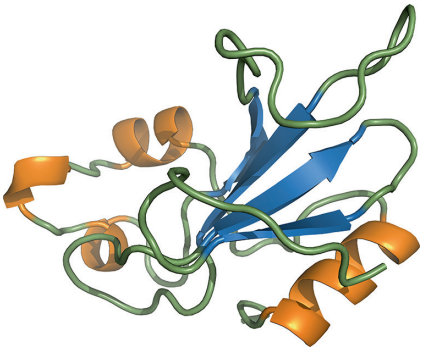

Abb. 1.18 Die Ribonuklease (1A2P) gehört zu den *alpha plus beta folds*. Eine einzelne Helix schmiegt sich gegen ein antiparalleles Faltblatt.

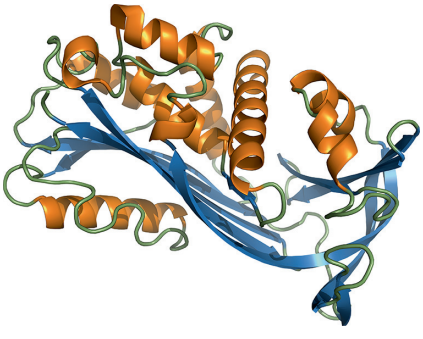

Abb. 1.19 Dieser Hydrolaseinhibitor (1HLE) ist eines der einfachsten Multidomänenproteine. Diese Faltungstypen enthalten jeweils mehrere Domänen, die zu unterschiedlichen Klassen gehören.

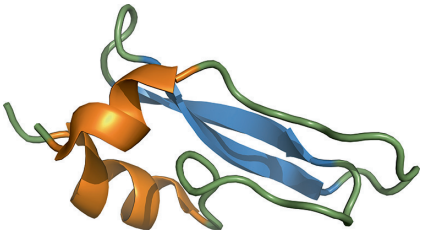

Abb. 1.20 Beispiel für ein kleines Protein. Dieser Hydrolaseinhibitor (1G6X) weist einen BPTI-ähnlichen Faltungstyp auf und wird als disulfidreicher *alpha plus beta fold* klassifiziert.

drei paarweise Kombinationen, die zur Unterscheidung von Proteinstrukturen genutzt werden können: Dies sind α mit α, α mit β und β mit β.

SCOP-Klassen Die SCOP-Klasse *all-alpha* wird von kleinen Proteinen dominiert. Häufig bilden die Helices ein auf und ab verlaufendes Bündel. Die Wechselwir-

kungen zwischen den Residuen der Helices sind nicht so präzise zu identifizieren wie bei β-Strängen, sodass eine genaue Klassifikation schwierig ist. Die *all-beta*-Proteine werden häufig aufgrund der Anzahl von β-Strängen feiner klassifiziert. Die Struktur der β-Stränge ist weniger starr als die von α-Helices, daher ist die Topologie der β-Faltblätter häufig gestört und es treten Verdrehungen auf. α-β Proteine können grob in solche Proteine aufgeteilt werden, die ein alternierend wechselndes Arrangement von α-Helices und β-Strängen längs der Sequenz aufweisen und solche, die eher isoliert liegende Sekundärstrukturen besitzen. Die erste Klasse schließt einige große und sehr reguläre Sekundärstrukturelemente ein, bei denen ein zentrales β-Faltblatt oder parallele β-Stränge auf beiden Seiten von α-Helices bedeckt werden. Die Abb. 1.15–1.20 zeigen typische Vertreter für diese Proteinklassen, die der SCOP-Datenbank entnommen wurden. Es ist in Klammern jeweils der PDB-Code angegeben, unter dem der Datensatz in der Strukturdatenbank PDB zu finden ist. Eine weitere Klasse bilden die Membranproteine. Typische Vertreter sind im Kapitel zur bioinformatischen Bearbeitung von Membranproteinen gezeigt.

1.16 Enzyme

Die interessanteste und wohl wichtigste Proteinklasse stellen die *Enzyme*. Sie wirken als Biokatalysatoren, d. h., sie beschleunigen biochemische Reaktionen. Hierbei werden *Substrate* meist in einer Kavität des Enzyms, dem *aktiven Zentrum*, gebunden und in *Edukte* umgesetzt. Bei den effizientesten Enzymen wie

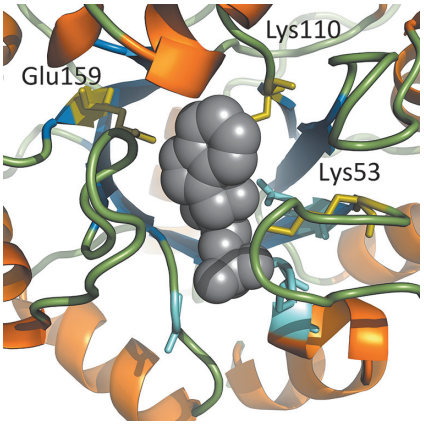

Abb. 1.21 Reaktionszentrum des Enzyms Indol-3-Glycerolphosphat-Synthase (TrpC, 1A53). Das Strukturgerüst des Enzyms ist wiederum abstrahiert, das Produkt IGP ist in Art eines Kalottenmodells grau dargestellt. Die an der Katalyse unmittelbar beteiligten drei Aminosäuren Lys53, Lys110 und Glu159 sind gelb, die drei an der Substratbindung beteiligten Residuen sind als hellblaue Stäbchen dargestellt.

der Triosephosphatisomerase [33] ist die Stoffumsetzung nur durch die Diffusionsgeschwindigkeit der Substrate und Edukte limitiert. Das oben erwähnte Rubisco hingegen schafft in der lebenden Zelle nur circa fünf Reaktionszyklen pro Sekunde [28] und gehört damit zu den langsamsten Biokatalysatoren. Die meisten Enzyme setzen sehr spezifisch genau ein Substrat um, weil nur dieses so im aktiven Zentrum zu liegen kommt, dass die Enzymreaktion ablaufen kann. An der Katalyse selbst sind häufig nur wenige Aminosäuren beteiligt. Auch für die räumlich korrekte Bindung der Substrate sind meist nur einige Aminosäuren verantwortlich. Die weiteren Aminosäurereste des Proteins sind beispielsweise dazu da, die für die Funktion wichtigen Reste korrekt zu positionieren, Bindetaschen geeigneter Größe auszubilden, die Stabilität des Proteins sicherzustellen, durch Bewegungen Signale zu übertragen, oder mit Residuen anderer Proteine zu wechselwirken. In der Abb. 1.21 ist das Reaktionszentrum des Enzyms Indol-3-Glycerolphosphat-Synthase dargestellt, das an der Tryptophansynthese beteiligt ist [34]. In der Abbildung sind die wenigen, direkt für die Katalyse wichtigen Residuen hervorgehoben.

Der Aufbau dieses Reaktionszentrums ist prototypisch und illustriert einige wichtige Eigenschaften von Enzymen:

- An der Stoffumsetzung selbst sind nur wenige Residuen beteiligt.
- Die lokale Umgebung dieser katalytischen Residuen bestimmt maßgeblich deren Orientierung und Beweglichkeit und andere chemische Eigenschaften wie die Ladungsverteilung im Reaktionszentrum.
- Sind prinzipiell mehrere, chemisch ähnliche Moleküle katalytisch umsetzbar, so ist neben anderen Kriterien die Größe der Bindungstasche ein wichtiger Parameter, der über die Prozessierung der Substrate entscheidet.

Diese Beobachtungen haben, wie wir später sehen werden, entscheidenden Einfluss auf das Design von bioinformatischen Algorithmen, mit denen die Funktion von Enzymen vorhergesagt werden soll.

1.17 Proteinkomplexe

Viele Proteine – und damit auch die Enzyme – erfüllen ihre Funktion nicht als einzelnes Protein (Monomer), sondern als Teil eines größeren Proteinkomplexes. Die einzelnen Elemente des Komplexes sind in der Regel nicht durch Atombindungen (d. h. kovalente Bindungen) miteinander verknüpft, sondern durch einfacher lösbare Wasserstoff- und Salzbrücken. Die Stärke des Zusammenhalts wird folglich durch die Größe des Protein-Protein-Interfaces und die Anzahl dieser nicht kovalenten Bindungen determiniert. Ein großer Komplex aus Proteinen und RNA-Molekülen ist das Ribosom. Das bereits erwähnte Rubisco lagert sich zu einem Komplex zusammen, der aus 16 Untereinheiten besteht. Häufig werden auch Komplexe beobachtet, die aus nur zwei Untereinheiten bestehen. Sind die Unter-

einheiten identisch, liegt ein *Homodimer* vor, sind sie unterschiedlich, so handelt es sich um ein *Heterodimer*.

In der Abb. 1.22 ist die als Heterotetramer vorkommende Tryptophansynthase gezeigt. Sie besteht aus je zwei Untereinheiten TrpA und TrpB und katalysiert die zwei letzten Schritte der Tryptophanbiosynthese. Die Tryptophansynthase besitzt einige typische Eigenschaften von Enzymkomplexen:

- Die Untereinheiten aktivieren sich gegenseitig, d. h. ihre Aktivität erhöht sich bei Komplexbildung.
- In diesem Komplex existiert ein hydrophober Tunnel, der eine Substratpassage vom aktiven Zentrum in TrpA hin zum aktiven Zentrum in TrpB ermöglicht und einen Verlust des Substrats durch Diffusion reduziert.
- Die Substratbindung induziert den Austausch sogenannter *allosterischer Signale*, die einen Einfluss auf die Katalyse haben. Der Transfer dieser Signale geht einher mit Konformationsänderungen von einzelnen Aminosäureseitenketten und ganzen Schleifen.

Dieses Beispiel macht deutlich, dass Proteine keine starren Objekte sind, sondern unterschiedliche Konformationen einnehmen, um z. B. Substrate in das katalytische Zentrum aufzunehmen.

Im rechten Teil der Abb. 1.22 sind die beiden Untereinheiten in Form von Kalottenmodellen dargestellt. Hierbei wird jedes Atom durch eine Kugel repräsentiert. Atomgrößen, Bindungswinkel und Bindungslängen entsprechen den physikalisch-chemischen Verhältnissen. Kalottenmodelle vermitteln ein realistisches Bild von der Packungsdichte und der Oberfläche der Proteine, während die Cartoon-Modelle besser geeignet sind, den Faltungstyp darzustellen.

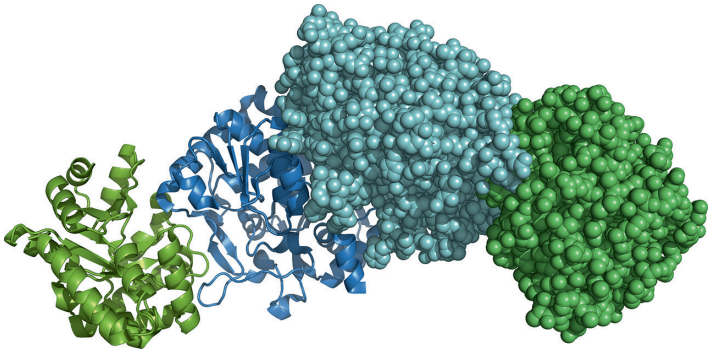

Abb. 1.22 Schematische Darstellung der Tryptophansynthase (2RHG). Dieses Enzym besteht aus zwei TrpA (grün) und zwei TrpB Untereinheiten (blau), die sich in einem Tetramer zusammenlagern. Links sind die Untereinheiten im Cartoon-Modus, rechts als Kalottenmodell dargestellt.

1.18 Fachbegriffe

In den folgenden Kapiteln sind wir auf biologische Fachbegriffe angewiesen. Die Wichtigsten, sofern nicht anderweitig im Text erläutert, werden hier kurz zusammengefasst und erläutert.

Die Begriffe homolog, ortholog und paralog, die Verwandtschaftsbeziehungen beschreiben, benötigen wir im Kontext von Genen und Genomen.

Homologe, orthologe, paraloge Gene Zwei Gene sind homolog, wenn sie beide von einem *gemeinsamen Vorfahren* abstammen. Diese Definition schließt orthologe und paraloge Gene mit ein.

Ortholog sind Gene aus *unterschiedlichen* Spezies, die sich durch Artenbildung aus einem gemeinsamen Vorfahren entwickelt haben.

Paralog sind Gene, die *im selben Genom* zu finden und durch Genduplikation entstanden sind.

Aus diesen Definitionen folgt, dass es keine graduelle Abstufung der Homologie gibt. Die Aussage, „zwei Gene oder Proteine sind zu x % homolog" ist falsch. Ihre Sequenzen mögen zu x % identisch oder ähnlich sein; aufgrund ihrer Abstammung sind sie jedoch entweder homolog oder nicht homolog.

Genotyp Der Genotyp ist die Summe der Gene, die in einem Genom vorkommen.

Phänotyp Der Phänotyp ist das äußere Erscheinungsbild einer Art. In der Genetik wird aus dem Vergleich unterschiedlicher Phänotypen auf die Funktion von Genen geschlossen.

Prokaryont Die Prokaryonten (auch Prokaryoten) sind diejenigen Arten, die keinen Zellkern besitzen. Dazu gehören die Bakterien und die Archaeen. Bakterien und Archaeen bilden nach gültiger Lehrmeinung jeweils eigene taxonomische Reiche.

Eukaryont Die Eukaryonten (oder Eukaryoten) sind diejenigen Arten, die einen Zellkern besitzen.

Mikroorganismen Als Mikroorganismen werden diejenigen Arten zusammengefasst, die mit dem bloßen Auge nicht zu erkennen sind. Dazu gehören Bakterien, Archaeen aber auch Pilze wie die Hefe *Saccharomyces cerevisiae*.

Gramfärbung Mit dieser Färbemethode können Bakterien aufgrund des Aufbaus ihrer Zellmembran in zwei große Gruppen eingeteilt werden. Diese werden grampositive bzw. gramnegative Bakterien genannt.

Genom Die komplette Erbinformation eines Lebewesens heißt Genom.

Metagenom Es wird angenommen, dass nur 1 % aller Mikroorganismen im Labor kultivierbar ist. Die Metagenomik versucht, die Gesamtheit aller Genome eines Biotopes zu bestimmen. Hierzu wird dem Biotop eine Probe entnommen, es wird DNA isoliert und deren Sequenz bestimmt. Die Menge der gefundenen DNA-Sequenzen nennt man Metagenom.

Systembiologie Die *Systembiologie* versucht, Organismen als Ganzes zu verstehen. Deswegen ist sie auf die Analyse des Zusammenwirkens vieler Gene oder

Proteine angewiesen. Zu den wichtigsten Werkzeugen der Systembiologie gehören *Hochdurchsatzmethoden,* die mit jedem Experiment umfangreiche Sätze von Messwerten erheben. Hochdurchsatzmethoden und ihre Anwendungen werden häufig im Kontext biochemischer Spezialdisziplinen genannt, deren Namen die Endsilbe „omik" tragen. Diese widmen sich dem Studium biologischer „Datensätze" deren Namen auf „om" enden. Zu den wichtigsten Disziplinen gehören *Genomik*, *Transkriptomik*, *Proteomik* und *Metabolomik*.

Genomik *Genomik* fokussiert sich auf die Erforschung des Genoms, d. h. die Gesamtheit aller Gene. Untersucht wird das Zusammenwirken der Gene, ihre Bedeutung für das Wachstum und die Entwicklung sowie für die Steuerung biologischer Systeme. Im Rahmen von Genomprojekten muss die Gesamtsequenz der DNA aufgeklärt und annotiert werden. Annotation ist der Prozess, in dem möglichst alle funktionstragenden Elemente identifiziert und hinsichtlich ihrer Funktion genau beschrieben werden. Hierfür werden bevorzugt bioinformatische Verfahren eingesetzt.

Transkriptomik *Transkriptomik* ist der Versuch, spezifische Expressionsmuster von Genen zu identifizieren und zu analysieren. Das *Transkriptom* ist das transkriptionelle Profil einer Zelle in einem spezifischen Zustand. Es wird aus der Menge biochemisch nachweisbarer mRNA-Moleküle abgeleitet. Dieser Ansatz beruht auf einem zentralen Dogma der Genombiologie. Es besagt, dass die Transkription von Genen genau dann erfolgt, wenn die zugehörigen Genprodukte aufgrund einer spezifischen Situation benötigt werden. Daher erlaubt der Vergleich von mRNA-Konzentrationen diejenigen Gene zu identifizieren, die unter den, durch die jeweiligen Proben repräsentierten, Bedingungen aktiviert werden. Allerdings reflektiert der mRNA-Status nicht den Proteinstatus einer Zelle. Der Grund für unterschiedliche mRNA und Proteinkonzentrationen sind die verschiedenen Abbauraten.

Proteomik *Proteomik* zielt darauf ab, Proteinkonzentrationen direkt zu bestimmen, um auf diese Weise einen exakten Status aktiver Genfunktionen abzuleiten. Dies ist eine heroische Aufgabe: Viele Proteine werden posttranslational modifiziert, sodass z. B. eine menschliche Zelle mehr als eine Million unterschiedlicher Proteinvarianten enthalten kann. Es ist sehr schwer, diese mit biochemischen Methoden zu unterscheiden.

Metabolomik *Metabolomik* beschäftigt sich mit dem Problem, all die Moleküle (die *Metaboliten*) zu identifizieren, die zu einem definierten Zeitpunkt in einer Zelle vorhanden sind. Zu dieser Menge gehören jedoch nicht DNA- oder RNA-Moleküle und auch nicht Enzyme oder Strukturelemente der Zelle.

Interaktives Arbeiten

Den Einsatz von Dotplots und die Berechnung paarweiser Alignments können mithilfe der Lernmodule geübt werden, die auf der begleitenden Website angeboten werden.

Literatur

1 Osawa, S., Jukes, T.H., Watanabe, K. und Muto, A. (1992) Recent evidence for evolution of the genetic code. *Microbiol. Rev.*, **56**, 229–264.

2 Jimenez-Montano, M.A. (1994) On the syntactic structure and redundancy distribution of the genetic code. *Biosystems*, **32**, 11–23.

3 Kagawa, Y., Nojima, H., Nukiwa, N., Ishizuka, M., Nakajima, T., Yasuhara, T., Tanaka, T. und Oshima, T. (1984) High guanine plus cytosine content in the third letter of codons of an extreme thermophile. DNA sequence of the isopropylmalate dehydrogenase of *Thermus thermophilus*. *J. Biol. Chem.*, **259**, 2956–2960.

4 Bernardi, G. und Bernardi, G. (1986) Compositional constraints and genome evolution. *J. Mol. Evol.*, **24**, 1–11.

5 Hori, H. und Osawa, S. (1987) Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. *Mol. Biol. Evol.*, **4**, 445–472.

6 Cox, E.C. und Yanofsky, C. (1967) Altered base ratios in the DNA of an *Escherichia coli* mutator strain. *Proc. Natl. Acad. Sci. USA*, **58**, 1895–1902.

7 Nghiem, Y., Cabrera, M., Cupples, C.G. und Miller, J.H. (1988) The mutY gene: a mutator locus in *Escherichia coli* that generates G.C-T.A transversions. *Proc. Natl. Acad. Sci. USA*, **85**, 2709–2713.

8 Galtier, N. und Lobry, J.R. (1997) Relationships between genomic G + C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.*, **44**, 632–636.

9 Salzberg, S.L., Delcher, A.L., Kasif, S. und White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucl. Acids Res.*, **26**, 544–548.

10 Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H. und Wright, F. (1988) Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity, *Nucl. Acids Res.*, **16**, 8207–8211.

11 Andersson, S.G. und Kurland, C.G. (1990) Codon preferences in free-living microorganisms. *Microbiol. Rev.*, **54**, 198–210.

12 Karlin, S. und Mrazek, J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.*, **182**, 5238–5250.

13 Sharp, P.M. und Li, W.H. (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.*, **24**, 28–38.

14 Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.*, **146**, 1–21.

15 Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13–34.

16 Sørensen, M.A., Kurland, C.G. und Pedersen, S. (1989) Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.*, **207**, 365–377.

17 Najafabadi, H.S. und Salavati, R. (2008) Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biol.*, **9**, R87.

18 Hawley, D.K. und McClure, W.R. (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucl. Acids Res.*, **11**, 2237–2255.

19 Birney, E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

20 Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.

21 Whelan, S. und Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.

22 Taylor, W.R. (1986) The classification of amino acid conservation. *J. Theor. Biol.*, **119**, 205–218.

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

24 Ponder, J.W. und 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.

25 Ramachandran, G.N., Ramakrishnan, C. und Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.

26 Linderstrøm-Lang, K.U. (1952) *Proteins and Enzymes*, Standford Univ. Press, Stanford.

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

28 Tabita, F.R., Hanson, T.E., Li, H., Satagopan, S., Singh, J. und Chan, S. (2007) Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. *Microbiol. Mol. Biol. Rev.*, **71**, 576–599.

29 Wierenga, R.K. (2001) The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Letters*, **492**, 193–198.

30 Sterner, R. und Höcker, B. (2005) Catalytic versatility, stability, and evolution of the $(\beta\alpha)_8$-barrel enzyme fold. *Chem. Rev.*, **105**, 4038–4055.

31 Knippers, R. (1995) *Molekulare Genetik*, G. Thieme, Heidelberg.

32 Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C. und Murzin, A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.*, **36**, D419–425.

33 Albery, W.J. und Knowles, J.R. (1976) Evolution of enzyme function and the development of catalytic efficiency. *Biochemistry*, **15**, 5631–5640.

34 Hennig, M., Darimont, B.D., Jansonius, J.N. und Kirschner, K. (2002) The catalytic mechanism of indole-3-glycerol phosphate synthase: crystal structures of complexes of the enzyme from *Sulfolobus solfataricus* with substrate analogue, substrate, and product. *J. Mol. Biol*, **319**, 757–766.