1
Biologische Grundlagen

In den folgenden Kapiteln beschäftigen wir uns meist mit Algorithmen, die Eigenschaften von Makromolekülen bewerten oder vergleichen. Für das Verständnis der Algorithmen und der zugrunde liegenden informatischen Methoden und Modellieransätze benötigen wir relativ wenige biologische Grundkenntnisse, die in diesem Kapitel eingeführt werden. Zu den wichtigsten molekularbiologischen Substanzklassen gehören DNA, RNA und Proteine. Dies sind Makromoleküle, die jeweils aus einer Abfolge kleinerer Bausteine bestehen. Die DNA ist beispielsweise aus Nukleotiden aufgebaut und deren lineare Anordnung kann in Form einer Zeichenkette (Sequenz) angegeben werden. Das Konzept der Sequenzen betrachten wir im nächsten Kapitel genauer, im Folgenden konzentrieren wir uns zunächst auf biochemische und biophysikalische Eigenschaften der genannten Molekülklassen.

Die *DNA* ist der wichtigste Datenträger in der Molekularbiologie; das Genom einer Spezies, das die komplette genetische Information enthält, ist in DNA-Molekülen codiert. In den letzten Jahrzehnten wurden Hochdurchsatzmethoden entwickelt, die es erlauben, DNA-Sequenzen mit geringem Aufwand und in kurzer Zeit zu ermitteln. Aus diesen Gründen werden bevorzugt Genomsequenzen bestimmt, da deren Kenntnis häufig ausreicht, die Komposition der anderen Makromoleküle (RNA und Proteine) abzuleiten. Die biologische Bedeutung der *RNA* hat durch neuere Erkenntnisse enorm zugenommen. Es ist klar geworden, dass RNA-Moleküle nicht nur an der Umsetzung der genetischen Information in Proteine beteiligt sind. Sie übernehmen in erheblichem Ausmaß auch Regulationsaufgaben, was lange unbekannt war. *Proteine* sind die wichtigsten Baustoffe aller biologischen Zellen. Sie geben den Zellen oft ihre Struktur und sind beispielsweise in Form von Enzymen essenzielle Komponenten der meisten Stoffwechselvorgänge.

Die *In-vivo*-Funktion von DNA, RNA und Proteinen kann nur anhand der dreidimensionalen Molekülstruktur verstanden werden. Im Vergleich zu den eher uniformen Raumstrukturen der DNA- und RNA-Moleküle bilden Proteine eine enorme Vielfalt unterschiedlichster Strukturen aus. Deswegen nimmt im Folgenden die Darstellung von Proteinarchitekturen einen breiteren Raum ein. Nach der Beschreibung typischer Protein-3-D-Strukturen beschäftigen wir uns mit Proteineigenschaften, die in bioinformatischen Algorithmen von Bedeutung sind.

Die in der Natur vorkommende Vielfalt von Lebewesen ist entstanden, weil sich vererbte organische Strukturen aufgrund von Evolutionsvorgängen in den biologi-

Bioinformatik – Grundlagen, Algorithmen, Anwendungen, 4. Auflage, Rainer Merkl.

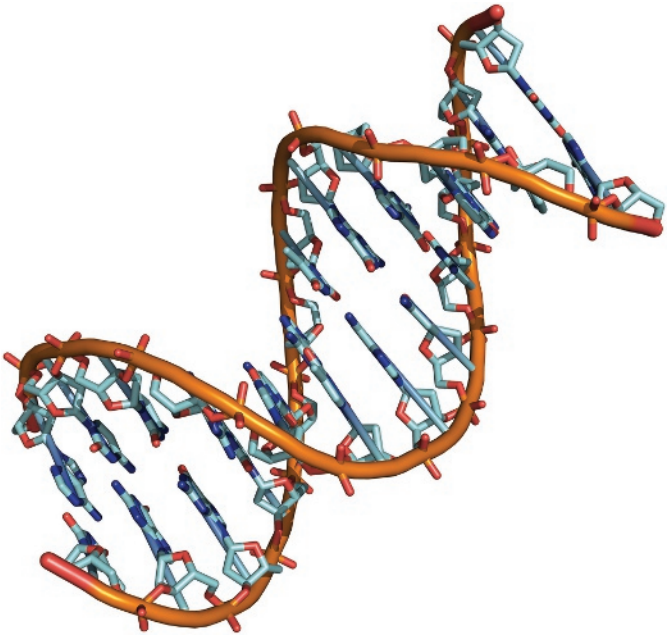


Abb. 1.1 Raumstruktur der DNA. In diesem DNA-Fragment ist die Doppelhelix gut zu erkennen. Die basischen Anteile der Nukleotide sind nach innen gerichtet und durch Wasserstoffbrücken verknüpft. Die Wasserstoffbrücken sind in dieser Abbildung nicht markiert. Außen verlaufen die Zucker-Phosphat-Anteile der polymerisierten Nukleotide; sie sind orange dargestellt. Die Raumstruktur dieses DNA-Fragments wurde mit experimentellen Methoden bestimmt, sodass die exakte Position aller DNA-Elemente bekannt ist und visualisiert werden kann.

schen Arten unterschiedlich entwickelt haben. Wir müssen uns daher auch mit dem Wesen von biologischen Evolutionsprozessen beschäftigen, da diese einen wesentlichen Teil der informatischen Modellbildung ausmachen. Das Kapitel schließt mit einer Definition wichtiger Fachbegriffe.

1.1 DNA

Im bioinformatischen Kontext beschreiben Sequenzen in der Regel eine bestimmte Abfolge von Einzelbausteinen, die aus einer kleinen und definierten Menge stammen. So sind DNA-Sequenzen einfache Modelle für Makromoleküle der Desoxyribonucleinsäure (abgekürzt DNS oder DNA), die in der Natur als fädige Struktur vorliegt. Die Grundbausteine sind vier Nukleotide, diese bestehen jeweils aus

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

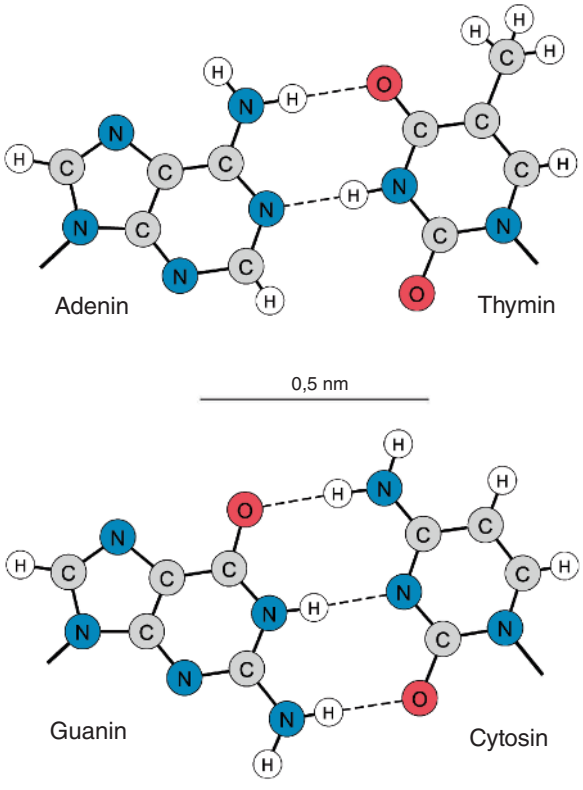


Abb. 1.2 Basenpaarungen in der DNA. In der als Doppelhelix bekannten DNA-Struktur liegen sich jeweils paarweise die Basen Adenin und Thymin oder Guanin und Cytosin gegenüber. Zwischen A:T-Paaren können zwei und zwischen G:C-Paaren drei Wasserstoffbrücken ausgebildet werden. Je höher der Anteil von G:C-Paaren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

Ein DNA-Strang ist aus einer Abfolge von Nukleotiden aufgebaut und in der Zelle kommt die 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). 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.

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 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-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 zum Bestimmen der Sequenz des Gegenstranges das *reverse Komplement* gebildet werden.

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 Nukleotidgruppe 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 Mi-

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 Tab. 2.2 in Kap. 2).

		2									
		T		**C**		**A**		**G**			
1	**T**	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	**T**	**3**
		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**	

Quelle: Basierend auf [1].

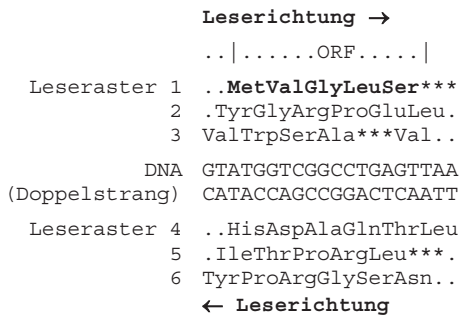


Abb. 1.3 Übersetzen eines DNA-Fragments in Proteinsequenzen. DNA kann in sechs Leserastern in Codonen übersetzt werden; pro Leserichtung ergeben sich jeweils drei Leseraster. In dieser DNA-Sequenz kommt nur ein ORF vor, die resultierende Proteinsequenz ist durch Fettdruck hervorgehoben. Ein ORF ist eine DNA-Teilsequenz, die durch ein Start- und ein Stoppcodon flankiert wird. Die Namen der Aminosäuren sind im Dreibuchstabencode angegeben, *** steht für die Sequenz von Stoppcodonen.

tochondrien, *Mycoplasma* und einigen Protozoen [1]. Stoppcodonen terminieren die für Proteine codierenden Gensequenzen.

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. Das Übersetzten der Gensequenz in eine Proteinsequenz beginnt an einem Startcodon und endet am nächsten Stoppcodon. Die Codonsequenz der drei Stoppcodonen ist eindeutig definiert, als Startcodon dient häufig „ATG“, das aber auch für die Aminosäure Methionin codiert.

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 Stoppcodonen auf oder es fehlt ein Startcodon. Nur ein kleiner Teil der ORFs codiert für Gene und diese bestehen in der Regel aus mehr als 80 Codonen.

Informationsgehalt der Basenpositionen

Der Informationsgehalt I der drei Basenpositionen im Codon ist unterschiedlich, es gilt $I(\text{Position 2}) > I(\text{Position 1}) > I(\text{Position 3})$ [2]. Für diese Ungleichheit ist der genetische Code verantwortlich: Ein Blick auf Tab. 1.1 macht klar, dass eine Mutation der dritten Base im Codon die Aminosäurenkomposition häufig nicht verändert. Zudem resultiert eine Mutation in der ersten Basenposition im Einbau einer Ami-

nosäure 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 und Cytosin an der DNA ist eine charakteristische Größe eines Genoms. Der mittlere GC-Gehalt von bakteriellen Genomen schwankt zwischen 20 und 75 % [3]. In G:C-Basenpaaren werden drei Wasserstoffbrückenbindungen ausgebildet, in A:T-Basenpaaren nur zwei; daher wurde lange vermutet, dass ein hoher GC-Gehalt des Genoms z. B. für thermophile [4] oder halophile [5] 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 [6]. 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 *Escherichia coli* wurde geschlossen, dass der GC-Gehalt vom DNA-Replikationssystem moduliert wird [1]. Bestimmte Mutationen im *mutT*-Gen von *Escherichia coli* induzieren Transversionen von A:T- nach G:C-Basenpaaren [7] und Mutationen im *mutY*-Gen Transversionen von G:C- nach A:T-Basenpaaren [8]. Die Genprodukte beider Gene sind an der DNA-Replikation oder DNA-Reparatur beteiligt. Neben dem mittleren GC-Gehalt eines Genoms sind auch lokale Schwankungen von Interesse: Der GC-Gehalt des menschlichen Genoms beträgt circa 42 %; es gibt aber sogenannte *CpG-Inseln*, in denen der GC-Gehalt mehr als 50 % beträgt. Da CpG-Inseln in der Nähe von Promotoren (siehe Abschn. 1.3) gehäuft vorkommen, wurden bioinformatische Verfahren entwickelt, um sie zu identifizieren [9]. Auch der GC-Gehalt von RNA-Molekülen wird untersucht, da es definierte Bereiche gibt, deren GC-Gehalt auf die optimale Wachstumstemperatur schließen lässt [10].

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 ähnlich. Codonhäufigkeiten des Bakteriums *Bacillus subtilis*, das zu beiden eine große phylogenetische Distanz aufweist, sind auffällig anders. Solche Unterschiede in den Codonpräferenzen erlauben es, die taxonomische Herkunft der DNA einzuschränken [11]. Statistische Verfah-

ren wie Markov-Ketten werden z. B. im Programm Glimmer [12] dazu genutzt, die Lage von Genen vorherzusagen. Hierfür wird die Zusammensetzung von ORFs mit der speziesspezifischen Codonhäufigkeit verglichen.

Synonyme Codonen

Der genetische Code wird als *degeneriert* bezeichnet, da einige Aminosäuren durch mehrere Codonen codiert werden. Codonen, die für dieselbe Aminosäure codieren, werden *synonyme Codonen* genannt. Wie Tab. 1.2 belegt, können sich die speziesspezifischen Häufigkeiten, mit denen synonyme Codonen verwendet werden, deutlich unterscheiden. In Korrelation mit den ungleichmäßigen Codonhäufigkeiten treten Unterschiede in den speziesspezifischen Konzentrationen von Transfer-RNA (tRNA)-Molekülen auf [13, 14]. Die tRNA ist an der Translation, also an der RNA-instruierten Proteinsynthese, beteiligt.

Bevorzugte Codonen

Bei manchen Spezies variieren Codonhäufigkeiten zudem stark zwischen einzelnen Genen [15]. In bestimmten Genen tritt speziesspezifisch eine Teilmenge der Codonen bevorzugt auf; Übersichten finden sich in [16, 17]. Diese Verzerrung der Codonhäufigkeiten (*codon usage bias*) ist positiv korreliert mit dem Grad der Genexpression, d. h. der Menge des synthetisierten Genprodukts [18]. Mögliche Ursachen für diese Verzerrung der Codonhäufigkeiten sind die bereits erwähnten Unterschiede in den tRNA-Konzentrationen, das Aufrechterhalten der maximalen Elongationsrate, die Kosten für das Korrekturlesen sowie unterschiedliche Translationsraten der Codonen [19]. Diese Verzerrung der Codonhäufigkeiten wird als „Strategie" inter-

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**	**3**
		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**	

pretiert, die Wachstumsraten zu optimieren [20, 21]. 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 [22]. Ganz generell illustrieren die erwähnten Befunde die komplexe Komposition codierender DNA-Sequenzen.

Codonhäufigkeiten 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; deswegen wird auch die Produktion dieser mRNA so 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 dieser 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 durch die bevorzugte Bindung von RNA-Polymerase an Erkennungsstellen (*Promotoren*), die unmittelbar vor Genen liegen, erreicht. 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
Aus einem Vergleich der Promotoren von *Escherichia coli* kann ein kanonischer Promotor gebildet werden. Dieser besitzt folgende Auffälligkeiten:

- 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.

Die Abb. 1.4 zeigt einen idealisierten Promotor; bekannte Promotoren weichen von dieser Sequenz mehr oder weniger stark ab.

Abb. 1.4 Konsensussequenz von *Escherichia coli*-Promotoren. Der untere der beiden DNA-Stränge wird ab Position +1 transkribiert; basierend auf [24].

Funktion von Transkriptionsfaktoren

Für die Einleitung der Transkription ist es notwendig, dass Transkriptionsfaktoren an den Promotor oder an zusätzliche Bindestellen wie *Enhancer* oder *Silencer* 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, die in *ein* mRNA-Molekül transkribiert werden. Die gemeinsam synthetisierten Genprodukte sind meist Elemente einer größeren Funktionseinheit oder tragen zur selben Stoffwechselleistung bei. So finden sich die bakteriellen Gene, die an der Biosynthese der Aminosäure Tryptophan beteiligt sind, in einem Operon. Das Identifizieren von Promotoren mit bioinformatischen Methoden hilft, Operons mit höherer Sicherheit vorherzusagen. Der Vergleich der Zusammensetzung von Operons gleicher Funktion, die aus unterschiedlichen Organismen stammen, kann dazu beitragen, die Funktion unbekannter Gene aufzuklären [23].

1.4 RNA

Bei höheren Eukaryonten ist nur für einen kleinen Bruchteil des Genoms die genaue Funktion bekannt [25]. 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 lange 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 erst langsam geklärt wird [26]. Diese Ergebnisse belegen für das Genom des Menschen [27] und der Maus, dass der größte Teil transkribiert wird. ncRNAs werden in kleine interferierende RNAs, mikro-RNAs 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 [28]. 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.

Im Vergleich zu Genen und Proteinen sind ncRNAs bislang weniger gut untersucht und ihre Raumstrukturen sind kaum bekannt. Allerdings werden mittlerweile mithilfe von Hochdurchsatzmethoden großen Datensätze generiert; parallel dazu entstehen neue bioinformatische Algorithmen mit dem Ziel, das Verständnis der RNA-Regulation zu fördern.

1.5 Proteine

Proteine sind ebenfalls lineare Makromoleküle; Sonderfälle, die vom linearen Aufbau abweichen, sind für uns nicht von Belang. Bausteine sind 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). Die Form und die chemische Beschaffenheit dieses Restes beeinflussen die Wechselwirkungen zwischen den Bausteinen. Die wichtigsten Wechselwirkungen sind Wasserstoffbrückenbindungen zwischen polaren Seitenketten.

Natur der Aminosäuren

Aufgrund des unterschiedlichen Aufbaus ihrer 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*, also solche, die innerhalb des Moleküls eine unterschiedliche Ladungsverteilung aufweisen. 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.

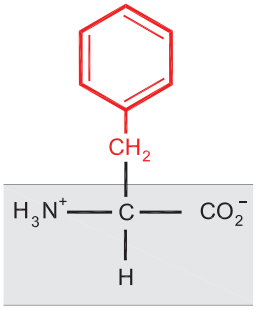


Abb. 1.5 Strukturformel der Aminosäure Phenylalanin. Der in allen Aminosäuren gleichartige Anteil ist in der Strukturformel grau unterlegt. In jeder Aminosäure ist mit dem zentralen C-Atom ein Wasserstoffatom (unten), eine Aminogruppe (links), eine Carboxylgruppe (rechts) und eine Seitengruppe (oben) 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 der Zusammensetzung der SWISS-PROT-Datenbank ermittelt. 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 [%]	8,4	5,5	4,0	5,4	1,3	3,9	6,8	7,1	2,2	5,9	9,8	5,9	2,4	3,8	4,7	6,7	5,3	1,1	2,9	6,9

Quelle: Basierend auf [29].

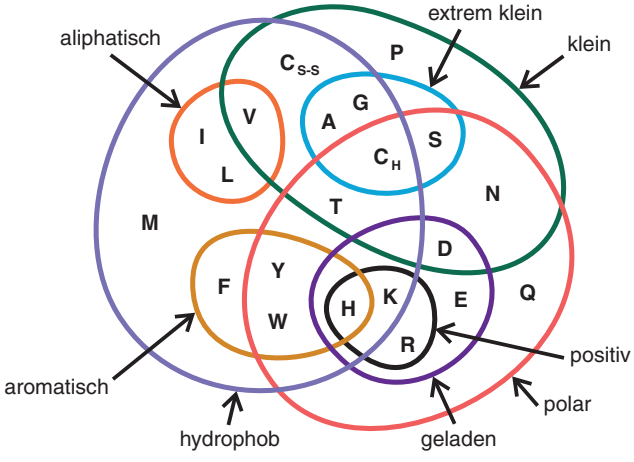


Abb. 1.6 Venn-Diagramm der 20 natürlichen, in Proteinen vorkommenden Aminosäuren. Die Aminosäuren wurden aufgrund der physikalisch-chemischen Eigenschaften gruppiert, die für die Tertiärstruktur von Proteinen wichtig sind. Die Aminosäuren sind im Wesentlichen in zwei Gruppen (polar und hydrophob) eingeteilt, 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. Der Einbuchstabencode wird im folgenden Kapitel genauer erläutert und ist in der Legende zu Tab. 1.3 angegeben. Abbildung nach [30], Abbildung 3a (S. 210)/mit freundlicher Genehmigung von Elsevier.

Die in Abb. 1.6 dargestellten Verwandtschaftsbeziehungen aufgrund der physikalischen und chemischen Eigenschaften der Aminosäuren sind die Grundlage für viele Sequenzvergleichs- und Alignment-Verfahren. Für die optimale Ausrichtung werden Scoring-Matrizen benötigt, die aus dem Vergleich einer Vielzahl ähnlicher Proteine ermittelt wurden und gemeinsame Eigenschaften von Aminosäuren widerspiegeln. 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 Peptidbindung verknüpft. Jede Polypeptidkette beliebiger Länge hat ein freies Aminoende (N-Terminus) und ein freies Carboxylende (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 einer Peptidbindung beteiligten sechs Atome liegen jeweils starr in einer Ebene. Da jede Aminosäure an zwei Peptidbindung beteiligt ist, wird der *Hauptkettenverlauf* einer Polypeptidkette durch die Angabe von zwei Winkeln pro Residuum definiert. Die ϕ- und ψ-Winkel geben die Drehung der beiden am Hauptkettenverlauf beteiligten Bindungen des zentralen C_α-Atoms jeder Aminosäure an (siehe Abb. 1.7). Beide Winkel unterliegen weiteren Einschränkungen, die sich aus der Natur des jeweiligen Aminosäurerestes herleiten. Die Rigidität der Peptidbindung und die sterische Hinderung zwischen Atomen der Haupt- und Seitenketten tragen zur Stabilisierung der Proteinkonformation bei. Der Hauptkettenverlauf dient häufig dazu, Faltungstypen von Proteinen zu charakterisieren und zu vergleichen. Die Hauptkette wird oft auch als *Proteinrückgrat* bezeichnet und heißt im Englischen *backbone*. In Abb. 1.7 ist der Aminosäurerest abstrahiert dargestellt; das erste, hier nicht gezeigte Kohlenstoffatom, das im Rest R auf das C_α-Atom folgt, wird C_β-Atom genannt.

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 Menge bevorzugter *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 oft eine beschränkte Menge aller möglicher Seitenkettenkonformationen betrachtet, die *Rotamere* genannt werden. Diese sind in Bibliotheken zusammengefasst [31] 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

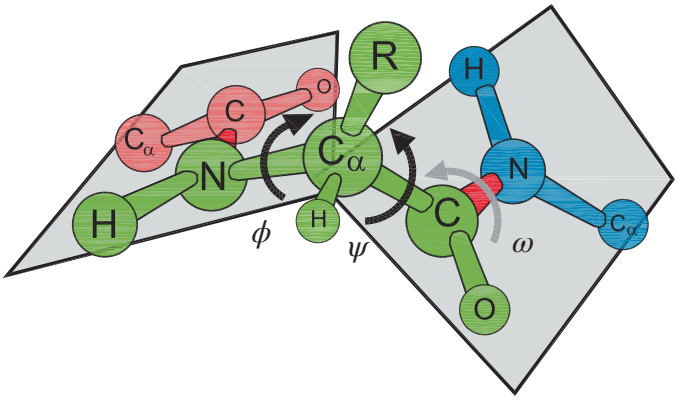


Abb. 1.7 Konformation der Peptidbindung. Die an einer Peptidbindung beteiligten sechs Atome liegen jeweils in einer Ebene, die grau markiert ist. In der Abbildung sind zwei derartige Bindungen gezeigt und rot markiert. Der hier grün markierte Aminosäurerest ist mit R bezeichnet. Die räumliche Anordnung des Hauptkettenverlaufs eines Polypeptids $\ldots -C_\alpha-C-N-C_\alpha- \ldots$ wird durch Paare von Winkeln (ϕ, ψ), die für jede Residuenposition anzugeben sind, bestimmt. Mit diesem Paar von Winkeln ist die Lage der durch die Peptidbindungen aufgespannten Flächen relativ zum C_α-Atom festgelegt. Der mit ω bezeichnete Winkel kann nur die Werte $+180°$ oder $-180°$ annehmen und die hier gezeigte Ausprägung ist die wesentlich häufigere.

Bindung resultieren jeweils 81 Rotamere. Beispiele für Rotamere werden in Abb. 1.8 vorgestellt. Die Menge der bekannten Proteinstrukturen erlaubt es, mit statistischen Methoden die Rotamerverteilungen in Abhängigkeit von den ϕ- und ψ-Winkeln der Hauptkette zu bestimmen. Solch hauptkettenspezifische (*backbone dependent*) Bibliotheken [32] 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, 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 Glycinresten. Der Einbau von Glycin bewirkt eine scharfe Wendung des Hauptkettenverlaufs. Diese Darstellung der Winkelkombinationen wird nach ihrem Entwickler *Ramachandran-Plot* genannt. Die erwähn-

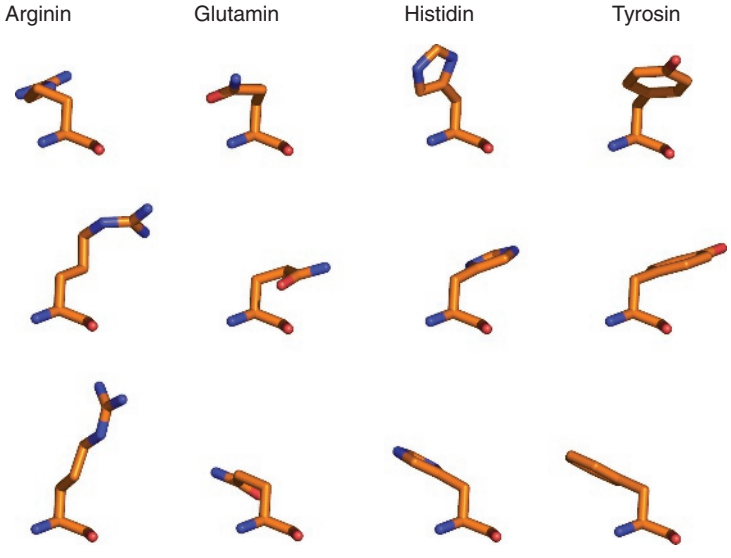


Abb. 1.8 Beispiele für Rotamerausprägungen. Rotamere sind in Proteinen häufig vorkommende Seitenkettenkonformationen. In der Abbildung sind für die Aminosäuren Arginin, Glutamin, Histidin und Tyrosin jeweils drei Rotamere gezeigt. Die Seitenkette von Arginin enthält vier drehbare Bindungen mit jeweils 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.

ten Sekundärstrukturelemente α-Helix und β-Faltblatt werden im folgenden Text genauer beschrieben.

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 [33]. Dies sind:

- Die *Primärstruktur*: Sie wird durch die Abfolge der Aminosäuren als Sequenz fixiert.
- Die *Sekundärstruktur*: Aus der Polypeptidkette falten sich Sekundärstrukturelemente, die regelmäßige Arrangements des Hauptkettenverlaufs ergeben.
- Die *Tertiärstruktur*: Sie beschreibt die räumliche Anordnung aller Atome eines Proteins im Raum.
- Die *Quartärstruktur*: Sie definiert die Anordnung von Proteinen in Proteinkomplexen.

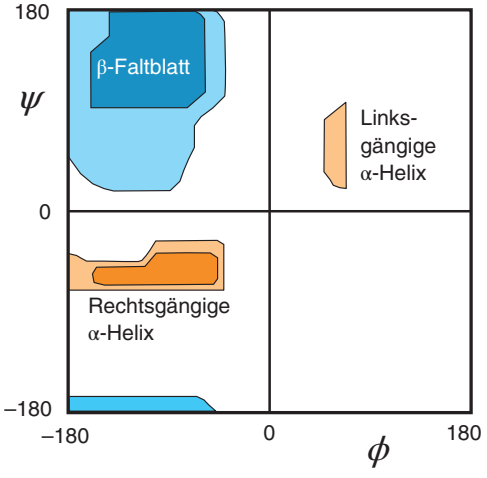


Abb. 1.9 Ramachandran-Plot. Je nach Zugehörigkeit zu einem Sekundärstrukturelement ergeben sich für die ϕ- und ψ-Winkel der Residuen charakteristische Kombinationen. Aufgrund energetisch ungünstiger Wechselwirkungen sind viele Winkelkombinationen selten oder sie kommen in Proteinen nicht vor.

Wir werden Algorithmen kennenlernen, die darauf abzielen, die 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 3-D-Teilstrukturen des Hauptkettenverlaufs einer Peptidkette. Bei der Klassifizierung von Sekundärstrukturelementen werden Art und Anordnung der Aminosäurereste ignoriert. Die Stabilisierung der Sekundärstruktur erfolgt über Wasserstoffbrückenbindungen zwischen den Imino- und Carbonylgruppen, *die zur Hauptkette gehören*. Zusätzlich zu diesen Bindungskräften wird die 3-D-Struktur eines Proteins im Wesentlichen durch schwache, nicht kovalente Wechselwirkungen der Aminosäureseitenketten, insbesondere durch Wasserstoffbrückenbindungen zwischen polaren Resten bestimmt. Diese Wechselwirkungen spielen aber 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 aufeinanderfolgender Residuen konstant, so ergeben sich helikale Strukturen. Unter diesen kommt die α-Helix am häufigsten vor. In der α-Helix besteht jeweils eine Wasserstoffbrückenbindung zwischen der C=O-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 fünf bis zehn 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:

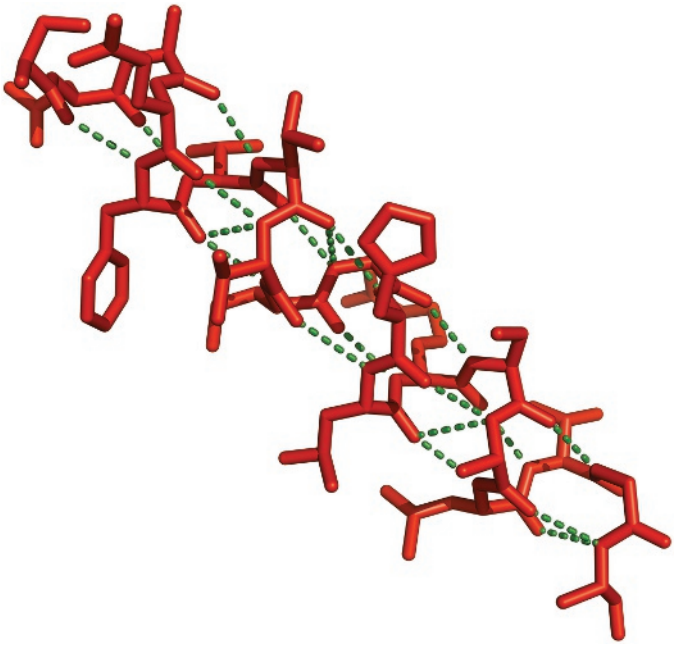


Abb. 1.10 Typische α-Helix. Wasserstoffbrücken, die zwischen Atomen des Proteinrückgrats ausgebildet werden, sind grün gestrichelt gezeichnet. Die Struktur ist als Stäbchenmodell dargestellt.

- *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 Seite 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 (βα)-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 Biosynthese der Aminosäure Histidin beteiligt. In HisF sind die acht (βα)-Einheiten durch weitere Sekundärstrukturelemente ergänzt. Das auf der Erde vermutlich mengenmäßig häufigste Protein ist das Enzym Rubisco. Es ist an der Fotosynthese beteiligt und besitzt ebenfalls eine $(\beta\alpha)_8$-artige Struktur [34]. Ausführlich wird diese Faltungstopologie in [35, 36] beschrieben.

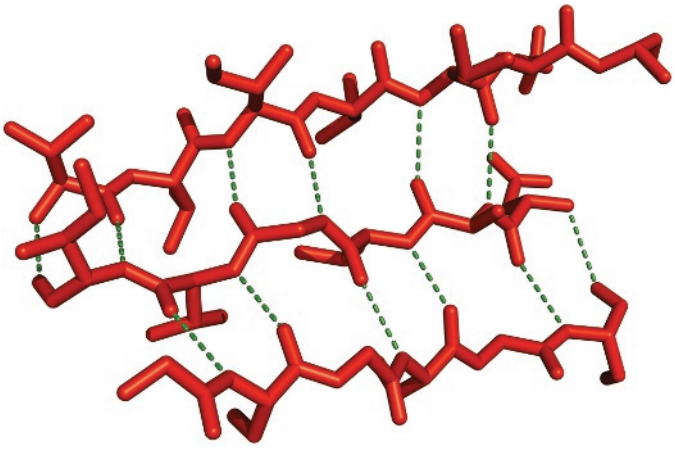


Abb. 1.11 β-Faltblatt bestehend aus drei Strängen. Wasserstoffbrücken zwischen Atomen der Hauptkette sind grün gestrichelt eingezeichnet. Die Struktur ist als Stäbchenmodell dargestellt.

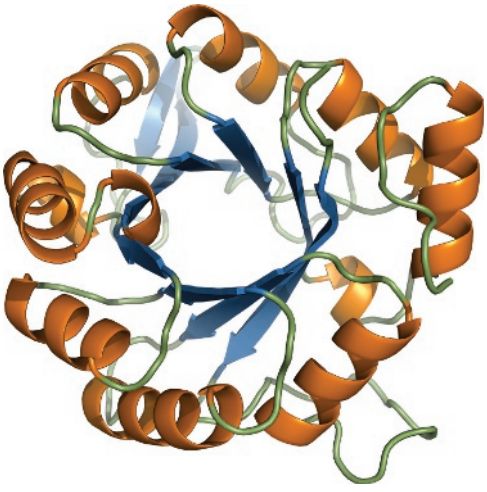


Abb. 1.12 Das $(\beta\alpha)_8$-Fass-Protein HisF. Beim Faltungstyp der $(\beta\alpha)_8$-Fässer bilden acht β-Stränge ein zentrales, in sich geschlossenes Faltblatt, das von acht α-Helices umgeben ist. Diese idealisierte Struktur ist häufig durch zusätzlich Schleifen oder andere Sekundärstrukturelemente 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) mit solchen Regionen, die keinerlei Ähnlichkeit zur Vergleichssequenz haben, ab. 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 3-D-Struktur eines CAP-Monomers dargestellt, das aus zwei Domänen besteht:

- 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, also Aggregate von zwei Monomeren, sind in Bakterien Transkriptionsaktivatoren, die mit mehr als 100 Promotoren wechselwirken [37].

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

Abb. 1.13 3-D-Struktur eines CAP-Monomers. Die N-terminale Domäne ist orange, die C-terminale Domäne türkis eingefärbt. *In vivo* lagern sich jeweils zwei CAP-Moleküle zu einem Dimer zusammen; basierend auf [37].

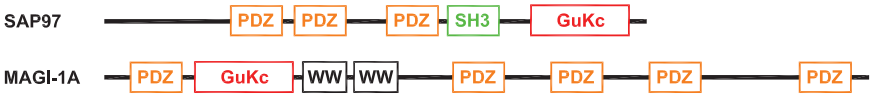


Abb. 1.14 Domänenstruktur des präsynaptischen Proteins SAP97 und des MAGI-1A Proteins.

Proteine SAP97 und MAGI-1A wie folgt beschreiben: Beide Proteine 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 in Abschn. 1.14 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 nicht mehr als circa 400 Faltungstypen gehören. Diese Faltungstypen werden jeweils durch eine Supersekundärstruktur charakterisiert und Proteine können

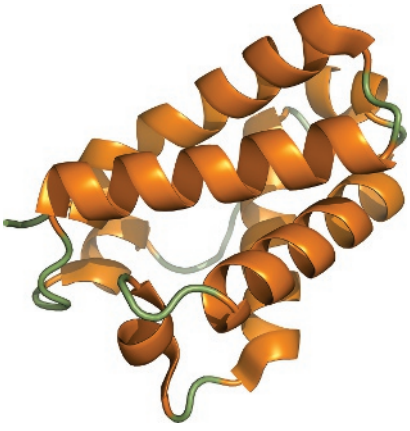


Abb. 1.15 Beispiel für ein *all-alpha*-Protein. Dieses Protein (PDB-ID 1DLW) besitzt einen globinähnlichen Faltungstyp. Die SCOP-Klassifikation lautet: sechs Helices, gefaltetes Blatt, teilweise geöffnet. In Klammern ist der Bezeichner angegeben, mit dem der Datensatz in der Strukturdatenbank PDB zu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

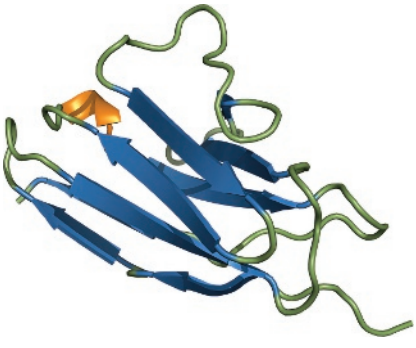


Abb. 1.16 Das Bence-Jones-Protein (PDB-ID 1BWW) ist ein *all-beta*-Protein. Die SCOP-Klassifikation lautet: Sandwich, sieben Stränge in zwei Faltblättern, einige Mitglieder dieses Faltungstyps besitzen zusätzliche β-Stränge.

aufgrund dieser Faltungstypen gruppiert werden. Im Kapitel zu Datenbanken (Abschn. 3.7) wird das Klassifikationssystem SCOP [38] vorgestellt, das auf einem solchen Schema beruht. Wie sehen repräsentative Vertreter der Faltungstypen aus? In den Abb. 1.15–1.19 werden Beispiele für die wichtigsten Faltungstypen im *Cartoon*-Modus präsentiert, hierbei wird auf die Wiedergabe der Seitenketten verzichtet. Diese Darstellung des Rückgrats vermittelt anschaulich 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, da Schleifen keine typischen Strukturmuster aufweisen. Aufgrund der Beschränkung auf zwei Klassifikationselemente existieren nur drei paarweise Kombinationen, die zur Unterscheidung von Proteinstrukturen genutzt werden: 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 Wechselwirkungen 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

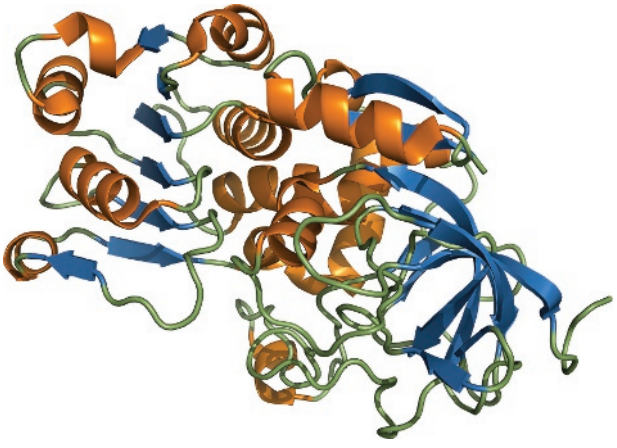


Abb. 1.17 Die NAD(P)-bindende Domäne des *Rossmann folds* (PDB-ID 2JHF) gehört zu den *alpha and beta folds*. Der Kern besteht aus drei Schichten, dazu kommt ein paralleles β-Faltblatt, bestehend aus sechs β-Strängen. Im Hauptkettenverlauf folgen α-Helices und β-Stränge unmittelbar aufeinander.

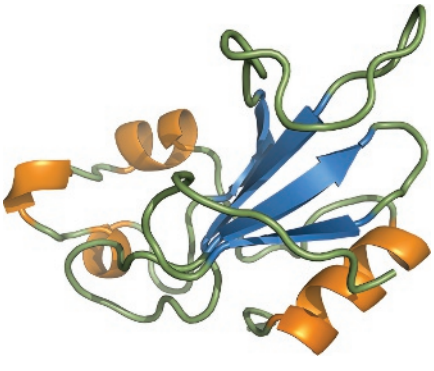


Abb. 1.18 Die Ribonuklease (PDB-ID 1A2P) gehört zu den *alpha plus beta folds*. Eine einzelne Helix schmiegt sich gegen ein antiparalleles Faltblatt. Die α-Helices und β-Stränge liegen getrennt voneinander in der Hauptkette.

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 voneinander getrennt 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. 1.15–1.19 zeigen typische Vertreter für Proteinklassen, die der SCOP-Datenbank entnommen wurden. Es ist in Klammern jeweils der Bezeichner angegeben, unter dem der Datensatz in der Strukturdatenbank PDB zu finden ist. SCOP klassifiziert Proteine auch aufgrund ihres *Typs*. Alle gezeigten Beispiele sind vom Typ *lösliche Proteine*. Daneben gibt es *intrinsisch ungeordnete* sowie *Faser-* und *Membranproteine*; auf Letztere wird im Kapitel zur bioinformatischen Bearbeitung von Membranproteinen (Kap. 21) eingegangen.

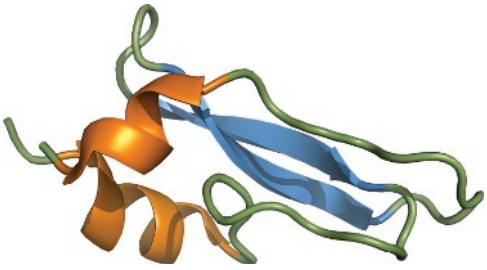


Abb. 1.19 Beispiel für ein kleines Protein. Dieser Hydrolaseinhibitor (PDB-ID 1G6X) weist einen BPTI-ähnlichen Faltungstyp auf und wird als disulfidreicher *alpha plus beta fold* klassifiziert.

1.16 Enzyme

Die interessantesten und funktionell wohl wichtigsten Proteine sind die *Enzyme*. Sie wirken als Biokatalysatoren und 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 der Triosephosphatisomerase [39] 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 [34] und gehört damit zu den langsamsten Biokatalysatoren. Die meisten Enzyme katalysieren sehr spezifisch eine Reaktion, weil nur bestimmte Substrate so im aktiven Zentrum zu liegen kommen, dass die Enzymreaktion ablaufen kann; es gibt allerdings Ausnahmen [40]. 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. 1.20 ist das Reaktionszentrum des Enzyms Indol-3-Glycerolphosphat-Synthase dargestellt, das an der Tryptophanbiosynthese beteiligt

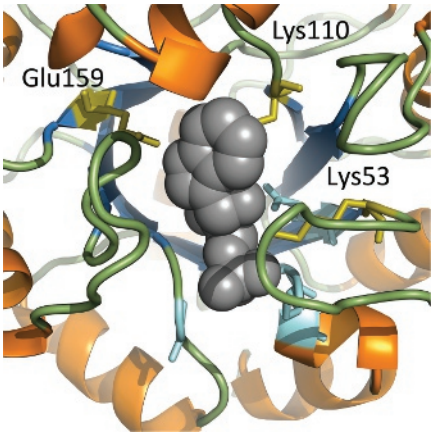


Abb. 1.20 Reaktionszentrum des Enzyms Indol-3-Glycerolphosphat-Synthase (TrpC, PDB-ID 1A53). Das Strukturgerüst des Enzyms ist wiederum abstrahiert, das Produkt IGP ist in der Art eines Kalottenmodells grau dargestellt. Die an der Katalyse unmittelbar beteiligten drei Aminosäuren Lys53, Lys110 und Glu159 sind gelb, die drei an der Substratbindung beteiligten Residuen sind als hellblaue Stäbchen dargestellt.

ist [41]. In der Abb. 1.20 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 miteinander verknüpft, sondern durch einfacher lösbare Wasserstoff- und Salzbrücken. Letztere sind eine Kombination aus einer Wasserstoffbrücke und einer ionischen Bindung. Die Stärke des Zusammenhalts wird folglich durch die Größe der Protein-Protein-Kontaktfläche (*interface*) und die Anzahl der darin vorkommenden, 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 Untereinheiten identisch, liegt ein *Homodimer* vor, sind sie unterschiedlich, so handelt es sich um ein *Heterodimer*.

In Abb. 1.21 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 von einer Untereinheit auf die andere geht mit Konformationsänderungen von einzelnen Aminosäureseitenketten und ganzen Schleifen einher.

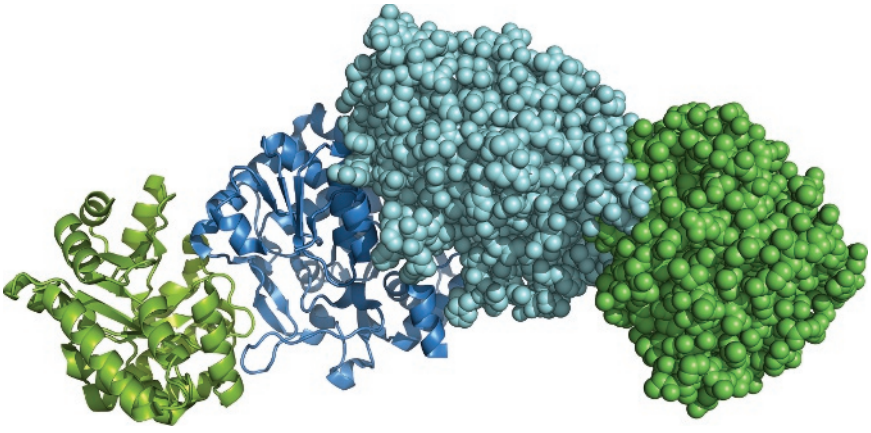


Abb. 1.21 Schematische Darstellung der Tryptophansynthase (PDB-ID 2RHG). Dieses Enzym besteht aus zwei TrpA- (grün) und zwei TrpB-Untereinheiten (blau), die sich in einem Tetramer zusammenlagern. Links sind die Untereinheiten als Cartoon-, rechts als Kalottenmodell dargestellt.

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 von Abb. 1.21 sind die beiden Untereinheiten der Tryptophansynthase 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 im linken Teil der Abbildung benutzten Cartoon-Modelle besser geeignet sind, den Faltungstyp zu zeigen.

1.18 Evolutionäre Prozesse

Die zuletzt vorgestellte Tryptophansynthase ist ein typischer Vertreter für einen Enzymkomplex, der in Bakterien und Archaeen, aber auch in Pflanzen vorkommt. In diesen phylogenetisch diversen Arten wird Tryptophan mithilfe weitgehend übereinstimmender Reaktionen synthetisiert. Aber bereits der Vergleich bakterieller *trp*-Operons, in denen üblicherweise auch die Gene für TrpA und TrpB enthalten sind, macht klar, dass sich deren Organisation und regulatorischen Mechanismen deutlich unterscheiden [42]. Ein Vergleich von TrpA- oder TrpB-Sequenzen lässt erkennen, dass sich selbst Sequenzen, die aus nahe verwandten Arten stammen, in Zusammensetzung und Länge deutlich unterscheiden können. Diese Variationen mögen zunächst verwundern, werden aber plausibel, wenn ihre evolutionäre Entstehung berücksichtigt wird.

Seit der Veröffentlichung der darwinschen Evolutionstheorie im Jahre 1858 gilt als gesichert, dass biologische Arten aufgrund eines ständigen Entwicklungsprozesses

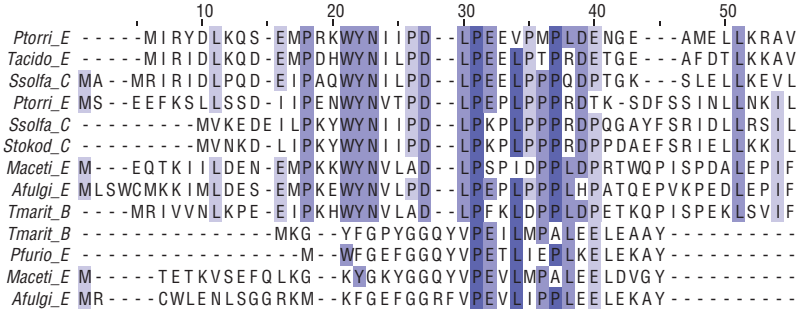


Abb. 1.22 Vergleich von TrpB-Sequenzen. Die abgekürzten Namen der Spezies, aus denen die Sequenzen stammen, enden mit einem B, wenn es sich um ein Bakterium handelt. Alle anderen Sequenzen stammen aus Archaeen: Ein endständiges E oder C gibt an, dass die Sequenz aus einem Euryarchaeon oder einem Crenarchaeon stammt. Die Aminosäuresequenzen sind im Einbuchstabencode abgegeben. Positionen im Protein, die mit identischen Aminosäuren besetzt sind, wurden durch einen blauen Hintergrund hervorgehoben. Bei der Ausrichtung der unterschiedlich langen Sequenzen wurden Lücken eingeführt, die durch Trennzeichen symbolisiert werden. Hier ist der Sequenzanfang gezeigt.

entstehen und sich kontinuierlich verändern. Die Eckpfeiler dieser Theorie sind die Folgenden:

- Alle Merkmale von Lebewesen sind im Genom codiert, das durch Fortpflanzung weitergegeben wird.
- Durch Veränderungen in der Basenzusammensetzung des Genoms (Mutationen) entstehen in jeder Generation Varianten, sodass sich einzelne Individuen in ihren Eigenschaften unterscheiden.
- Durch natürliche Selektion, z. B. durch sich verändernde Lebensräume oder durch zufällige Gendrift, werden einige Varianten bevorzugt an die nächste Generation weitergegeben.

Das Zusammenspiel zwischen Mutationen und natürlicher Selektion über viele Generationen hinweg erzeugt einen kontinuierlichen Prozess, mit dem auch die Entstehung neuer Arten durch Abspaltung erklärt werden kann. Häufig lässt sich, beispielsweise bei Proteinen, die Abstammung von einem gemeinsamen Vorfahren jedoch noch erkennen: Die Proteine, die als *homolog* bezeichnet werden, weisen oft noch eine deutliche Ähnlichkeit ihrer Sequenz oder Struktur auf. Andrerseits kann diese Art der Fortentwicklung dazu führen, dass die verbliebenen Ähnlichkeiten nur noch schwer zu detektieren sind.

In Abb. 1.22 sind 13 TrpB-Sequenzen, die aus Bakterien und Archaeen stammen, in einem Alignment zusammengefasst. Diese optimale Ausrichtung der Sequenzen macht zum einen klar, dass sie unterschiedlich lang sind: Deswegen mussten Lücken eingefügt werden. Zum anderen unterscheiden sich die Sequenzen deutlich in ihrer Zusammensetzung: Nur wenige Residuenpositionen sind mit identischen Aminosäureresten besetzt.

Dieses Beispiel belegt, dass einfache Vergleichsalgorithmen nicht ausreichen, das Evolutionsgeschehen exakt zu verfolgen:

> Die stochastische Komponente des Evolutionsprozesses erzwingt einen „gewissen“ Aufwand bioinformatischer Algorithmen.

Es folgt, dass die evolutionäre Entwicklung umso genauer modelliert werden muss, je präziser Funktion und Eigenschaften biologischer Objekte untersucht werden sollen. Diese Steigerung der Ansprüche wird uns in den restlichen Kapiteln permanent begleiten.

1.19 Fachbegriffe

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

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

Taxonomische Nomenklatur
Jede biologische Art hat einen binären Namen wie *Homo sapiens*, wobei *Homo* eine Gattung (Genus) benennt. Wichtige taxonomische Kategorien oberhalb der Gattung sind Familie, Ordnung, Klasse, Stamm (Phylum) und Reich. In dieser Reihenfolge nimmt der Verwandtschaftsgrad der subsumierten Arten ab.

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 und im anderen Fall 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 wird Metagenom genannt.

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. Ein 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.

In vivo*, *in vitro

Prozesse, die im lebendigen Organismus ablaufen werden als *in vivo* bezeichnet, solche, die unter Laborbedingen beobachtet werden, als *in vitro*.

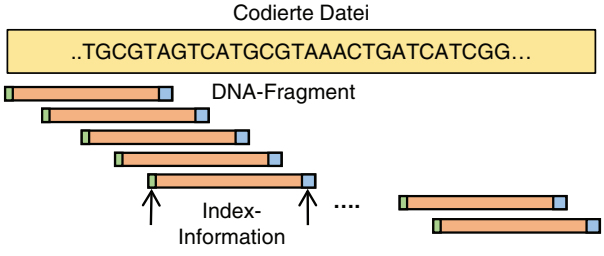
Exkurs: DNA als Langzeitspeicher, Shakespeare ersetzt Gene

Das sichere Speichern von Daten über lange Zeiträume ist nicht einfach. Dokumente mit hoher national- oder kunsthistorischer Bedeutung werden in Deutschland im Barbarastollen bei Oberried auf Mikro- und Farbfilm archiviert. Für das Filmmaterial wird erwartet, dass es für mindestens 500 Jahre ohne Informationsverlust lagerfähig ist. Technologien wie Magnetbänder, DVDs oder CD-ROMs bieten langfristig keinen sicheren Schutz vor Datenverlust. Es besteht also Bedarf, sich nach Alternativen umzusehen.

Alle lebenden Zellen enthalten einen Langzeitspeicher mit enormer Kapazität, die DNA. Nick Goldman, der am *European Bioinformatics Institute* (EBI) in Hinxton tätig ist, hat 2013 alle 154 Sonette von Shakespeare und einige weitere Texte und Bilder in DNA gespeichert [43]. Die Dateien machten insgesamt 739 kB Plattenkapazität aus und wurden in einem speziellen Code auf 153 335 DNA-Fragmente, die

117 Nukleotide lang waren, übertragen. Jedes Fragment enthielt zusätzlich einen Index, der die Datei und die Position innerhalb der Datei angab. Die Fragmente wurden mit einem Standard-DNA-Syntheseautomaten hergestellt, gefriergetrocknet und von den USA nach Deutschland verschickt. Nach der Sequenzierung am *European Molecular Biology Laboratory* (EMBL) in Heidelberg konnten die Dateien vollständig rekonstruiert werden.

In 1 g DNA können 215 PB gespeichert werden und DNA degradiert nicht, im Gegensatz zu technischen Datenträgern. Im Moment verhindern die Kosten die breite Anwendung dieser Technologie: Für das Synthetisieren von 2 MB Daten müssen circa 7000 € und für das Lesen circa 2000 € (Stand 2017) bezahlt werden [44].



Konzept für die Datenspeicherung in DNA, basierend auf [43].

Interaktives Arbeiten

Werkzeuge zur 3-D-Darstellung von DNA- und Proteinmolekülen sowie weiteres Übungsmaterial werden auf der begleitenden Website angeboten.

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 Hildebrand, F., Meyer, A. und Eyre-Walker, A. (2010). Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6: e1001107.

4 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.

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

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

7 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.

8 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.

9 Tahir, R.A., Zheng, D.A., Nazir, A. und Qing, H. (2019). A review of computational algorithms for CpG islands detection. *J. Biosci.* 44: 143.

10 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.

11 Novoa, E.M., Jungreis, I., Jaillon, O. und Kellis, M. (2019). Elucidation of codon usage signatures across the domains of life. *Mol. Biol. Evol.* 36: 2328–2339.

12 Delcher, A.L., Bratke, K.A., Powers, E.C. und Salzberg, S.L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23: 673–679.

13 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.

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

15 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.

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

17 Ermolaeva, M.D. (2001). Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.* 3: 91–97.

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

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

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

21 Novoa, E.M. und Ribas de Pouplana, L. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet.* 28: 574–581.

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

23 Zallot, R., Oberg, N. und Gerlt, J.A. (2019). The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* 58: 4169–4182.

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

25 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.

26 ENCODE Project Consortium, Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C. und Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.

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

28 Amaral, P.P., Dinger, M.E., Mercer, T.R. und Mattick, J.S. (2008). The eukaryotic genome as an RNA machine. *Science* 319: 1787–1789.

29 UniProt Consortium (2019). UniProt: A worldwide hub of protein knowledge. *Nucl. Acids Res.* 47: D506–D515.

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

31 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.

32 Shapovalov, M.V. und Dunbrack Jr., R.L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 19: 844–858.

33 Linderstrøm-Lang, K.V. (1952). *Proteins and Enzymes*. Stanford: Standford Univ. Press.

34 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.

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

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

37 Lawson, C.L., Swigon, D., Murakami, K.S., Darst, S.A., Berman, H.M. und Ebright, R.H. (2004). Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.* 14: 10–20.

38 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.

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

40 Peracchi, A. (2018). The limits of enzyme specificity and the evolution of metabolism. *Trends Biochem. Sci.* 43: 984–996.

41 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.

42 Merino, E., Jensen, R.A. und Yanofsky, C. (2008). Evolution of bacterial *trp* operons and their regulation. *Curr. Opin. Microbiol.* 11: 78–86.

43 Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B. und Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* 494: 77–80.

44 Service, R. (2017). DNA could store all of the world's data in one room, https://doi.org/10.1126/science.aal0852.