Notizen zum Abschnitt

Struktur einzelsträngiger Nukleinsäuren

der BC4-Vorlesung

Walter Fontana
Institut für Theoretische Chemie

Wien, Dezember 1996

Inhaltsverzeichnis

1 Einleitung

Viele RNAs liegen als Einzelstrang vor. Die Chargaff-Regeln ($\chi_A = \chi_U, \chi_G = \chi_C$) sind nicht erfüllt. Hinweise erhält man auch aus der Abhängigkeit hydrodynamischer und optischer Eigenschaften (Viskositätsuntersuchungen und Birefringenzmessungen) von rRNA. Die Viskosität von rRNA Lösungen variiert stark mit der Salzkonzentration. Eine Steigerung der Ionenstärke bewirkt bis zur 100-fachen Abnahme der (Grenz-)Viskosität. Die Viskosität von DNA (sowie homopolymeren Nukleinsäuren) wird dagegen kaum beeinflusst. Schluss: RNA ist ein kontraktiler Einzelstrang-Coil. Francis Crick schlägt bei einem Mittagessen in Wien vor, dass RNA unter bestimmten Bedingungen auf sich selbst zurückfalten kann und eine Sekundärstruktur einnimmt. Es wird vermutet, dass RNA eine partielle Helix-Struktur hat (40%-60%) und die Doppelhelix-Abschnitte von flexiblen Einzelstrang-Abschnitten ("loops") unterbrochen werden. Dies ist bis heute die Grundlage unseres Verständnisses der Sekundärstruktur von RNA.

Die Sequenz der tRNA-Ala wird von Holley 1965 vorgestellt und wurde die Grundlage für das Kleeblatt-Modell. In den späten 70er Jahren führten vergleichende Sequenzanalysen zu einem Vorschlag für die Sekundärstruktur von rRNA. Unser gegenwärtiges Verständnis der Tertiärstruktur von RNA beruht auf den röntgenkristallographischen Untersuchungen von tRNA (Kim 1974, Robertus 1974). Die Wechselwirkungen, die die 3D Struktur bedingen, sind neuer Art: Basen-Tripel und Wechselwrikungen zwischen Basen und Phosphodiester-"Rückgrat". Die Tertiärstruktur von rRNA ist noch nicht endgültig geklärt.

Wie verhalten sich Nukleinsäuren in wässriger Lösung? Welche Arten geordneter Strukturen nehmen sie ein? Wie kann man diese beobachten? Warum sind RNA und DNA strukturell so unterschiedlich, obwohl sie sich in ihrer Zusammensetzung doch nur im 2'-Sauerstoffatom unterscheiden? Was folgt daraus für unser Verständnis der Rolle dieser Polymere in der Organisation der Zellvorgänge? Welche Abstraktionen erweisen sich als nützlich in der Beschreibung von RNA Strukturen? Welche Methoden gibt es zur ihrer Berechnung?

Im Gegensatz zu Proteinen haben Nukleinsäuren eine wesentlich geringere Vielfalt funktionaler Gruppen und chemische assays zur Strukturuntersuchung sind daher kaum möglich. Man könnte annehmen, dass diese chemische

Einfalt sich letztlich auch in einer strukturellen Einfalt widerspiegelt. Das ist aber ganz und gar nicht der Fall. Einzelsträngige Nukleinsäuren weisen ein phantastisches Spektrum struktureller Vielfalt auf. Ein Verständnis dieser Vielfalt ist allerdings leichter als bei Proteinen. Nukleinsäuren bestehen im Wesentlichen aus nur vier Bausteinen. Es stehen daher entsprechend weniger fundamentale Wechselwirkungen zum Strukturaufbau zur Verfügung. Die meisten Wechselwirkungen, die die Struktur langer Nukleinsäuren bedingen, treten bereits bei den kleinsten Fragmenten auf: den Dinukleotiden oder wenig längeren Ketten. Wirkungsvolle experimentelle Techniken können deshalb auf kleine Fragmente angewandt werden. Von diesen ist der Schluss zu Eigenschaften längerer Ketten gestattet. Dies gelingt selbst dann, wenn die Struktur nicht nur von der Basenzusammensetzung, sondern von der Sequenz – der Reihenfolge der Basen in der linearen Kette – abhängt. Im Gegensatz dazu sind Dipeptide schlechte Modelle für die Struktur ganzer Proteinen.

Antworten auf Fragen nach der Wirkungsweise von Proteinen verlangen meistens eine Kenntnis der detaillierten 3D Konformation. Das ist ein schwieriges Unterfangen. Dagegen scheint es als könne man wichtige Teilaspekte der Funktion von Nukleinsäuren bereits auf einem viel abstrakterem Niveau - dem Niveau der Sekundärstruktur - verstehen. Die Sekundärstruktur von Nukleinsäuren entsteht dadurch, dass ein Einzelstrang durch Ausbildung von intramolekularen Basenpaaren (etwa der Watson-Crick Art) auf sich selbst zurückfalten kann. Unter der Sekundärstruktur einer RNA-Sequenz versteht man im Wesentlichen eine Liste der gepaarten Basen. Der Unterschied zu Proteinen besteht darin, dass die Basenpaar-Wechselwirkung binär ist – sie beschränkt sich überwiegend auf 2 Basen – und "digital" – entweder zwei Basen paaren oder sie tun's nicht. Eine solcher Strukturbegriff weist einen diskreten kombinatorischen Charakter auf, dessen Logik uns im letzten Teil dieses Abschnitts beschäftigen wird.

2 Struktur-Unterschiede RNA/DNA

Die überwiegende Mehrzahl von RNAs liegt als kovalenter Einzelstrang vor. RNA/DNA Hybrid-Doppelstränge treten als Zwischenprodukte bei der Transkription auf; entweder bei der normalen Transkription, in der ein RNA Strang von einer DNA Vorlage kopiert wird, oder bei der reversen Transkription, bei der anhand einer RNA-Matrize ein DNA Strang entsteht. Die

reverse Transkriptase benötigt zudem einen RNA primer um die DNA Synthese zu starten und daher tritt auch eine vorübergehende Situationen auf, in der ein RNA Abschnitt kovalent an einen DNA Abschnitt gebunden ist.

Warum findet man keine natürlich vorkommenden Nukleinsäure-Ketten, in denen Ribonukleotide und Deoxyribonukleotide gemischt auftreten? Der Grund dafür ist, dass Ribo- und Deoxyribo-Ketten *in vivo* Helixkonformationen mit recht unterschiedlicher Geometrie bevorzugen.

Erinnern wir uns zunächst an die helikalen Geometrien der DNA. Die A- und B-Formen (eigentlich sind es "Familien", da es eine Reihe von Varianten dieser Grundformen gibt) bestehen aus antiparallel laufenden Strängen, die rechtshändig gewunden sind. In der beigelegten Tabelle sind einige der charakteristischen Helix-Parameter gegenübergestellt. Der wesentliche geometrische Unterschied zwischen A und B Form ist (a) die Neigung der Paare relativ zur Längsachse der Helix (tilt) – in der A-Form sind sie stark geneigt, kaum jedoch in der B-Form – und der Abstand der Basenpaare von der Helix-Längsachse – er ist grösser bei der A-Form. DNA A- und B-Form können rasch durch Veränderung der Feuchtigkeit in den Fibern ineinander umgewandelt werden.

RNA-DNA Hybridhelices nehmen die A-Form an, selbst in wässriger Lösung. RNA Doppelhelices liegen ebenfalls und sogar ausschliesslich als Helices der A-Familie vor. Bei schwacher Ionenstärke hat die RNA A-Helix 11 Basen pro Umdrehung, ein tilt der Basenpaare zur Helixachse von $16 - 19°$ und eine Ganghöhe von 30 Å. De A'-Variante bei grösserer Ionenstärke hat 12 Basen pro vollständiger Drehung, ein tilt von 10° und eine Ganghöhe von 36 Å. Die Torsionswinkel entsprechen sonst der A-Familie wie wir sie von der DNA her kennen. Nie wurde bei RNA die B-Form beobachtet. Die Ursache dafür kann nur die Gegenwart der $2'$-Hydroxylgruppe sein. Diese hat einen Einfluss auf den pucker der Ribose. Das wird im Folgenden diskutiert.

Die Minimierung sterischer Konflikte zwischen den Ringsubstituenten führt dazu, dass der Ribose-pucker in Nukleotiden auf zwei Bereiche eingeschränkt ist: entweder der pucker ist am $C_{3'}$ lokalisiert und zwar *endo* (über dem Furanose-Ring liegend zum $C_{5'}$ hin orientiert) – die Pseudorotationsphase nimmt Werte zwischen 0° und 36° an – oder er sitzt am $C_{2'}$, ebenfalls *endo* – die Pseudorotationsphase ist zwischen 144° und 190°. Substituenten, die in der einen Phase axial liegen, liegen in der anderen äquatorial. Zwei Übergänge vom Süd-pucker ($C_{2'}$ *endo*) zum Nord-pucker ($C_{3'}$ *endo*) sind logisch möglich:

via O_4 *exo* oder via O_4 *endo*. Energetisch ist nur der *endo* Weg realisierbar. (Die Base und der C_5-Substituent sind entlang der *endo* Route äquatorial und behindern sich daher weniger.) Monomere Ribonukleotide interkonvertieren rasch zwischen beiden Formen, $C_{3'}$ *endo* und $C_{2'}$ *endo*, während monomere Deoxyribonukleotide den $C_{2'}$ *endo* pucker etwas bevorzugen. Elektronegative $2'$ Substituenten (Dipolmoment!) verschieben das Gleichgewicht in Richtung $C_{3'}$ *endo*.

In einer Helix sind die Konformationen der einzelnen Nukleotide natürlich nicht mehr voneinander unabhängig, sondern werden von den stacking Wechselwirkungen zwischen den Basen organisiert. In der Geometrie der A-Form ist der Ribose pucker ein $C_{3'}$ *endo*, während er in der B-Helix als $C_{2'}$ *endo* vorliegt (IUPAC-bürokratisch gesehen ist es die $C_{3'}$ *exo* Variante, die der $C_{2'}$ *endo* sehr ähnlich ist :-). RNA Dopelhelices liegen ausschliesslich in der A-Form vor. Im Unterschied zu den Mononukleotiden, ist der Ribose pucker im Kontext einer Helix also auf eine Konformation festgelegt. Hingegen kann im Fall der DNA-Doppelhelix der Deoxyribose pucker in beiden Konformationen vorliegen, was zur A- und B-Form führt. Die beiden Formen konvertieren je nach Wasseraktivität ohne dass Basenpaarungen aufgehen müssten. Der Grund warum RNA in der Helix-Struktur so konservativ ist, ist eigentlich noch immer nicht klar. Ein Trend zur Bevorzugung der $C_{3'}$ *endo* Konformation (A-Form) ist mit der Elektronegativität der $2'$ OH Gruppe im Einklang (s.o.), kaum aber eine Ausschliesslichkeit. Man vermutet dass die $2'$ OH Gruppe als H-Donor an Wasserstoffbrücken mit dem benachbarten Nukleotid beteiligt ist. Eine Möglichkeit wäre eine H-Brücke zum nächsten $O_{4'}$, oder eine H_2O vermittelte Wechselwirkung mit dem $3'$ Phosphat, oder aber auch einfache sterische Hinderung durch die Hydroxy-Gruppe. Die Gründe für die ausschliessliche A-Form in RNA Helices sind wahrscheinlich vielfältig und systemischer Natur, obwohl sie alle letztlich mit der Anwesenheit der $2'$-Hydroxygruppe zusammenhängen müssen.

Jetzt versteht man jedenfalls warum nicht Ribo- und Dexoxyribonukleotide in ein und demselben Polymer beliebig gemischt auftreten. Polyribonukleotid-Helices liegen in der A-Form vor, während unter physiologischen Bedingungen Polydeoxyribonukleotide die B-Form bevorzugen; wegen des stark unterschiedlichen Basen-Tilts in den A- und B-Helix Formen könnte die Kontinuität des Basen-"stackings" nicht aufrechterhalten werden. Somit entfiele die wesentliche Stabilisierung der gesamten Helixstruktur. Eine Koexistenz von längeren Abschnitten in A- und B-Form wäre denkbar: damit Basen-

stacking erhalten bleibt würde allerdings die Schnittstelle zwischen A- und B-Form einen Knick (kink) in der Helix verursachen. Ein Beispiel könnte der kovalente Hybridstrang sein, der als Zwischenzustand bei der RNA geprimten DNA Synthese auftritt.

Zusammenfassend nocheinmal der Steckbrief der standard helikalen Anordnungen:

1. Base ist *anti* relativ zum Zucker, d.h. Glycosyl ($C_{1'}$-N) Torsionswinkel ξ ist zwischen -90 und -160 Grad.
2. Zucker pucker ist $C_{2'}$ *endo* (Deoxyribose) oder $C_{3'}$ *endo* (Ribose).
3. $C_{4'}$-$C_{5'}$ Torsion γ im *+sc* Bereich ($\approx$ 60 Grad).
4. C-O Torsionswinkel β und ϵ sind *trans* ($\approx$ 180 Grad).
5. in rechtsdrehenden Helices sind beide P-O Torsionswinkel α und ξ im *-sc* Bereich ($\approx$ 300 Grad).

3 Struktur und Stabilität von Nukleinsäure-Einzelsträngen

Wir kennen nun die Doppelhelices von DNA und RNA, die durch Watson-Crick Basenpaarung entstehen. Wie sieht aber die Konformation einer Nukleotidkette aus, wenn sie als Einzelstrang vorliegt? Die Situation liegt beispielsweise bei der Transkription von DNA in RNA vor. Wenn eine solche Einzelkette Abschnitte enthält, die zueinander komplementär sind, dann kann sie intramolekulare Watson-Crick-Paare bilden. Mit diesem Fall werden wir uns ausführlich später beschäftigen. Zunächst betrachten wir Ketten, die keine komplementären Abschnitte enthalten, nämlich (synthetische) Homopolymere wie etwa poly-dA oder poly-A.

Sind solche Polymere in irgendeiner Weise geordnet? Es gibt eine Reihe möglicher Ordnungsstrukturen: die Basen können stacken und die Kette in einer Einzel-Helix-Konformation halten – wie eine Doppel-Helix aus der ein Strang entfernt worden wäre. Das andere Extrem entstünde, wenn die stacking

Wechselwirkung ohne Gegenstrang nicht gehalten werden kann und der Einzelstrang ein Zufalls-Knäuel wird. Eine weitere Möglichkeit sind Mehrfach-Helix-Konformationen, weil auch identische Nukleotide zueinander Wasserstoffbrücken ausbilden können. Diese sind nicht von der Watson-Crick Art, und die Helices könnten daher anders ausehen als im Fall der DNA. Es werden aber auch Konformationen zwischen vollständig geordneter Struktur und ungeordnetem Knäuel zu erwarten sein.

Das Ausmass an Ordnung hängt stark von Temperatur, pH, Ionenstärke und Lösungsmittel ab. Die wichtigsten Methoden, die Auskunft über die Struktur einzelsträngiger Polynukleotide in Lösung geben, sind die bei Polymeren üblichen: optische Methoden (wie NMR, CD) und hydrodynamische Methoden (wie Viskositätsmessungen). Wenn homopolymere Nukleinsäuren Zufallsknäuel wären, dürften sich die optischen Eigenschaften des Polymers nicht wesentlich von denen des Monomers unterscheiden. Bei hohen Temperaturen (100 C) ist dies für poly-A und poly-C auch der Fall. Bei tieferen Temperaturen kommt es zu starken Abweichungen, die allesamt auf die Anwesenheit zumindest abschnittsweiser Regelmässigkeiten in der Konformation schliessen lassen. So zeigen poly-A und poly-C bei neutralem pH und bei Raumtemperatur wesentlich stärkere CD Signale als die Monomeren. Die CD-Signale sind zwar in ihrer Intensität stark temperturabhängig, nicht aber in ihrer Form. Das bedeutet, dass welche geordneten Regionen auch immer existieren mögen, diese sich nicht in ihrer Art sondern nur in ihrer Häufigkeit verändern. Die wichtigste Bande im Absorptionsspektrum liegt bei ungefähr 260 nm. Gestackte Basenanordnungen zeigen dabei geringere Absorption als ungestackte – das ist als hypochromer Effekt bekannt. Dieser wird auch bei poly-A und poly-C beobachtet. Man kann also annehmen, dass die Ordnungsbereiche aus stacks von Basen bestehen und dass ihre Ausdehnung mit sinkender Temperatur zunimmt.

Man muss bedenken, dass solche Schlüsse nur anhand von wesentlich mehr Daten unterschiedlichster Art gezogen werden dürfen. So kann man allein anhand von CD und hypochromen Effekt nicht weitere Ordnungsmuster wie Basenpaarungen und Interkalation ausschliessen. Die weitgehende Unabhängigkeit der spektroskopischen Eigenschaften von der Ionenstärke erlauben im Fall von poly-C eine Interkalation (oder allgemein Strukturen in denen Ladungen sich nahe kommen würden, wie etwa Mehrfachhelices) auszuschliessen. IR-Daten schliessen hingegen eine Basenpaarung aus.

Viskosimetrische Messungen des Gyrationsradius von poly-C ergeben nur bei fast durchgehendem stacking (i.e., bei tiefen Temperaturen) eine drastische Zunahme des Gyrationsradius. Wir haben es also mit einer Kette zu tun, die bei tiefen Temperaturen eine Einfachhelix ist und mit Temperatur-Zunahme "schmilzt", d.h. die gestackten helikalen Abschnitte werden kürzer und sind entsprechend hüfiger von lokal ungeordneten Konformationen unterbrochen. Aus dem ersten Abschnitt wissen wir, dass die dihedralen Winkel im Ribosephosphodiester-Rückgrat starken sterischen Einschränkungen unterliegen. Dies verleiht der Kette eine gewisse Steifheit, sodass sie selbst als "Zufallsknäuel" noch einen hohen Gyrationsradius aufweist. Erst wenn die gestackten Abschnitte fast durchgehend sind, nimmt der Gyrationsradius rasch zu. Dieser Sachverhalt ergibt sich auch aus der statistischen Mechanik von Polymerkonformationen nach Flory.

Wie sieht die poly-C Helix aus? Sie gehört zur bekannten A-Familie von Helices, zu der auch die RNA Doppelhelix gehört. Poly-C hat aber nur 6 Monomere pro Drehung. Die Torsionswinkel im Helix-Rückgrat sind dabei den Torsionswinkeln der Mononukleotide viel ähnlicher als sie es in der RNA Doppelhelix sind. Soweit die kristallographische Struktur bei neutralem pH. Bei kleineren pH Werten wird das C am N-3 protoniert. Dabei wird ein C-C Basenpaar möglich, dass wahrscheinlich zwei poly-C Stränge in eine parallele Doppelhelix überführt. Ganz schlüssig sind die Daten aber nicht.

Ähnlich verhält es sich mit poly-A. Bei neutralem pH liegt eine Einfachhelix vor. Im sauren Bereich bildet sich dagegen eine Doppelhelix mit protoniertem A-A Paar. Die Einzel-Helix gehört wiederum zur A-Familie und hat neun Basen pro Drehung. Die Torsionswinkel sind wie im poly-C Fall dem Monomer sehr ähnlich. Der Zucker pucker ist $C_{3'}$ *endo* wie sich das für die A-Familie gehört. Das deoxy-Analogon poly-dA hat hingegen den pucker $C_{2'}$ *endo*, wie in der DNA Doppelhelixkonformation.

Zur Illustration der vielfältgien Ordnungsmuster bei homopolymeren Nukleinsäuren diene noch poly-dT und poly-U. Poly-dT weist eine rechts-drehende helikale Konformation auf in der die Basen nach "aussen" gedreht sind und nicht stacken. Die Stabilität ist entsprechend reduziert. Poly-U ist interessant weil es als Homopolymer bei niedriger Temperatur eine Sekundärstruktur ausbildet. Poly-U faltet auf sich selbst zurück und bildet eine doppelsträngige antiparallele Haarnadel mit asymmetrischen U-U Paar. Poly-dU (!) bildet keine Sekundärstruktur aus. Die Ursache dafür kann wiederum

nur die 2′ Hydroxygruppe sein. Wie sie aber genau wirkt ist unklar. Poly-G bildet schliesslich gar eine rechtshändige 4-fach Helix; ein Verhalten das bereits beim Monomer Guanosin sich ankündigt: Guanosin und etliche seiner Derivate aggregieren aus wässriger Lösung zu einem Gel mit tetrameren wasserstoffvernetzten Anordnungen.

4 Die Tertiärstruktur von RNA am Beispiel der tRNA

Das Interesse an einzelsträngigen Nukleinsäuren erschöpft sich nicht in der faszinierenden und schwierigen statistischen Mechanik rotationsmässig eingeschränkter langer Ketten. Das eigentliche Interesse gilt der starken Sequenzabhängigkeit ihrer Struktur. Wie bereits erwähnt, können einzelsträngige Nukleinsäuren intramolekulare Watson-Crick Paare ausbilden. Die Kette faltet dadurch auf sich selbst zurück und bildet helikale Abschnitte. Das führt notwendigerweise auch zu Schleifen (loops) verschiedenster Art und eröffnet den Nukleinsäuren eine enorme Vielfalt struktureller Formen. Dies wiederum ermöglicht spezifische Wechselwirkungen mit anderen Molekülklassen. Im letzten Jahrzehnt wurde deutlich dass Nukleinsäuren auch katalytisch aktiv sein können. Wir werden auf die Struktur-Funktions Beziehungen in RNA noch zzurückkommen. RNAs haben mit Proteinen ein reiches Repertoire an Strukturen und die Möglichkeit zu ihrer Variation (durch Variation der Sequenz) gemeinsam. Was aber RNA von Proteinen ganz entscheidend abhebt, ist die Kopierbarkeit der Sequenz: die Spezifizität der Basenpaarung, die für Struktur verantwortlich ist, ermöglicht gleichzeitig ihre Matrizenfunktion. RNA ist derzeit die einzige bekannten Objektklasse, die Genotyp und Phänotyp in einem einzigen Molekül vereinigt.

Wir wenden uns nun der tRNA und ihrer Tertiärstruktur zu. Die tRNA dient hier als *pars pro toto* um einige Prinzipien der sequenzabhängigen 3D-Struktur einzelsträngiger Nukleinsäuren zu illustrieren. Für Biophysiker war die tRNA die erste hochauflösende Röntgenstrukturanalyse einer RNA (Kim, 1974 und Robertus, 1974) und eines RNA/Protein Komplexes (Rould, 1989). tRNA forderte die Biochemiker zur ersten Sequenzbestimmung einer RNA heraus (Holley, 1965). Für Chemiker war tRNA die erste Synthese eines biologisch aktiven Gens (Ryan, 1979). tRNA spielte natürlich eine entschei-

dende Rolle bei der Aufklärung des genetischen Codes. Für Molekularbiologen wurde die Wechselwirkung zwischen tRNA und den Aminoacyltransferasen zum meist studierten Fall spezifischer RNA/Protein Wechselwirkungen. Molekular-Archaeologen nutzten tRNA zur Aufklärung evolutionärer Beziehungen zwischen Organismen und zur Datierung des genetischen Codes.

tRNA Moleküle gibt es in jeder Zelle. Die tRNA Population schwankt von 29 Spezies in *Mycoplasma* - dem kleinsten autonomen selbst-reproduzierendem Organismus - bis zu mehr als 150 Spezies in einer Säugerzelle. Die Hauptfunktion dieser RNA Klasse ist die Aktivierung von Aminosäuren zur Proteinbiosynthese. Für diese Funktion gibt es üblicherweise mehrere tRNA Spezies für jede Aminosäure (Isoakzeptoren). Zwei kritische Regionen der tRNA sind das Anticodon, das mit einem messenger RNA Codon wechselwirkt, und das Akzeptorende mit dem universellen CCA Abschnitt, dessen Adenosin an der $2'$ oder $3'$ Position mit der dieser tRNA spezifisch zugeordneten Aminosäure verestert wird. Unterschiedliche Zellkompartimente - Cytoplasma, Chloroplasten, Mitochondrien - enthalten ihre eigenen unterschiedlichen Proteinsynthese-Maschinerien; Chloroplasten und Säugermitochondrien haben eigene tRNAs, die sich von den zytoplasmatischen unterscheiden.

Man kennt derzeit etwa 2000 tRNA oder tRNA-Gen Sequenzen aus mehr als 200 Organismen und Organellen. Ihre Länge variiert zwischen 72 und 95 Nukleotiden; die meisten dieser Sequenzen lassen sich in die bekannte Kleeblatt-Sekundärstruktur falten. Eine bemerkenswerte Ausnahme sind die tRNAs von Säuger-Mitochondrien, von Flagellaten und Nematoden. Diese zeigen starke Abweichungen von der Standard Kleeblatt-Struktur. Eine weitere Eigenheit der tRNAs ist ihr hoher Gehalt an modifizierten Nukleosiden. Das Ausmass modifizierter Basen erreicht bis zu 25% in Human-tRNA und ist wesentlich geringer in bakterieller tRNA. Gegenwärtig sind die Strukturen 80 modifizierter Basen bekannt.

Die Biosynthese von tRNA geschieht durch Transkription von DNA. Es entsteht dabei eine längere precursor-tRNA. Ein einziges Transkript kann auch mehrere (21 in *B. subtilis*) Kopien verschiedener tRNAs enthalten. In Eukaryonten ist das Transkript monomer. Mehrere Exo- und Endonukleasen sind nötig um die überschüssigen Nukleotide zu entfernen. Einige der Basen werden direkt am Transkript methyliert, während andere in ihrer Originalversion ausgeschnitten und durch andersweitig synthetisierte und modifizierte Basen ersetzt werden.

In einigen Fällen enthält das tRNA Transkript Introns. Es ist bemerkenswert, dass tRNA Gene zwei unterschiedliche Arten von Introns aufweisen, die auch auf unterschiedliche Weise entfernt werden. In Eukaryonten werden die Introns (Länge 8-60) auf die übliche enzymatische Art ausgeschnitten, während einige Chloroplasten-Transkripte sich selbst "spleissen": ihr Intron (458 Nukleotide !) weist Sequenzähnlichkeiten mit den group I Introns auf. Evolutionär weit auseinanderliegende Bakterien weisen ebenfalls 200 Nukleotide "kurze" group I Introns in ihren tRNA Transkripten auf. tRNA wird auch editiert: an einigen Positionen werden die transkribierten Standardbasen in andere Standardbasen umgewandelt.

Nach diesem biochemischen Steckbrief der tRNA kehren wir zu ihrer Struktur zurück. Zunächst eine Klärung zum Begriff der Sekundärtsruktur. Ein Basenpaar teilt eine offene Region in zwei offene Regionen unterschiedlicher Grösse. Der entartete Fall, in dem eine Region die Grösse 0 hat, ist nichts anderes al ein Stack zweier aufeinanderfolgender Paare. Unter der Sekundärstruktur einer Nukleinsäure-Sequenz versteht man nun eine Liste von Basenpaaren mit der einzigen Einschränkung, dass sie keine Paare zwischen offenen Regionen enthält. Letztere nennt man Pseudoknoten und diese werden – wie alle anderen Konformationsaspekte – zu den Tertiärkontakten gezählt. Wir werden auf den formalen Begriff der Skundärstruktur später noch genauer eingehen. Er sei hier nur erwähnt, weil er nützlich in der Diskussion der Tertiärstruktur ist.

Die Kleeblatt-Sekundärstruktur besteht aus vier gepaarten Regionen. Diese werden Akzeptor-, Anticodon-, D- und T-Region oder -Arm (oder loop) genannt. In der Akzeptor-Region kommen die beiden $3'$ und $5'$ Enden zusammen. Das $3'$ Ende enthält die CCA Sequenz deren A mit der entsprechenden Aminosäure verestert wird. Die D- und T-Regionen heissen so, weil ihre zugehörigen loops immer ein Dihydrouridin (D) und ein Ribothymidin (T) enthalten. Die Längen der gepaarten Regionen und ihrer Schleifen ist im allgemeinen konstant. Die überschüssigen Nukleotide längerer tRNAs bilden den sogenannten variablen oder V-loop, der zwischen 4 und 21 Nukleotide enthalten kann. Wenn man tRNAs vergleicht, dann stellt man fest, dass verhältnismässig wenige Nukleotide invariant sind. Das wollen wir gleich festhalten, denn es sagt bereits etwas wichtiges über die Sequenz-Struktur Zuordnung aus: Sie ist nicht bijektiv. Es gibt viele verschiedene Sequenzen, die in die gleiche "Struktur" falten - zumindest auf der Ebene der Sekundärstruktur. Das Thema werden wir später wieder aufnehmen. Die invarianten Nukleotide

sind in der beigelegten Abbildung eingekreist und befinden sich vornehmlich in den loops der Sekundärstruktur. Man sieht auch dass einige dieser Invarianten an Pseudoknoten-Paarungen - also Tertiärkontakten - beteiligt sind.

Die Röntgenstrukturanalyse ergibt eine L-förmige Tertiärstruktur. Sie besteht aus zwei zueinander fast senkrecht liegenden Doppel-Helices. Eine schematische Darstellung dieser L-Form und ihres Zusammenhangs mit der Sekundärstruktur findet man ebenfalls in der Abbildung. Die Helices gehören erwartungsgemäss zur A-Familie. Die beiden Enden des "L" bestehen aus dem CCA Akzeptor-Ende und dem Anticodon. Akzeptor- und T-Arm sind an ihren inneren Enden gestackt und bilden somit eine durchgehende 11 Nukleotide lange Doppelhelix. Anticodon- und D-Arm sind ähnlich angeordnet, stacken aber nicht. Zwischen beiden helikalen Abschnitten ist ein Knick von 26 Grad. Die Tertiärtsruktur wird zusätzlich zu den Pseudoknoten von einer Reihe nicht-standard Basenpaaren und Basentripel in Form gehalten. Sie treten gehäuft in der Knickregion des "L"s auf.

Ein nicht-Watson-Crick Paar ist das G-U Paar. G-U Paare trifft man recht häufig im Inneren von RNA Helices an, und sie zählen bei der Ermittlung der Sekundärstruktur zum Standard-Paarungsrepertoire. Sie passen relativ gut in die Watson-Crick Geometrie der RNA Helix und erzeugen nur eine geringfügige Ausbuchtung des Phosphatrückgrates. Das G_{15}-C_{48} Paar ist ein "reverse Watson-Crick-Paar" und die gegenüberliegenden Kettenabschnitte verlaufen parallel. Ferner gibt es ein "reverse Hogsteen-Paar" (mit antiparalleler Ausrichtung) zwischen dem methylierten m^1A_{58} und dem T_{54}. Die Watson-Crick Anlagestelle ist beim m^1A_{58} durch Methylierung blockiert und Hogsteen-Paarung ist die einzige Alternative. Das G_{26} ist zweifach methyliert und paart ausgesprochen nicht-Watson-Crickisch mit A_{44}. Starke sterische Hinderung zwingt das Paar in eine nicht-planare Anordnung. Das Paar befindet sich genau an der Schnittstelle zwischen Anticodon-Arm und D-Arm. In der nicht-planaren Anordnung stackt das GA Paar mit beiden Helices und stabilisiert somit den 26-Grad Knick zwischen diesen Armen.

Es existieren auch eine Reihe von Basen-Tripel, die aus einem Watson-Crick Paar bestehen, das über eine oder zwei Wasserstoffbrücken mit einer weiteren Base verknüpft ist. Details findet man in der beigelegten Abbildung. Das A-A Basenpaar im A-U-A Tripel hat die gleiche Geometrie, wie man sie aus der protonierten poly-A Doppelhelix kennt, der wir im letzten Kapitel begegnet sind. In all diesen Tripel sind die Kettenverläufe notwendigerweise einmal

parallel und einmal antiparallel.

In Hefe $tRNA^{Phe}$ sind 42 von 76 Basen an helikalen Strukturen des standard A-Typs beteiligt. Trotzdem nehmen 71 Basen an stacking Wechselwirkungen teil. Die nicht-gestackten sind unter anderem das terminale A und die beiden Dihydrouracile an Position 16 und 17. Die beiden letzteren sind nicht planar und nicht aromatisch und mögen stacking daher gar nicht.

Eine weitere wichtige Spielart der stacking-Wechselwirkung ist in tRNA zu beobachten: Interkalation, d.h. das Einfügen einer Base zwischen zwei Aufeinanderfolgenden an einem benachbarten Strang. Insgesamt gibt es vier Interkalationen in der Hefe $tRNA^{Phe}$. Für diese Wechselwirkung muss der Abstand der benachbarten Basen, zwischen die interkaliert wird, etwas gestreckt werden. Dies geschieht über das Riboserückgrat indem der pucker Zustand an der Ribose eines Nukleotids vom standard $C_{3'}$ *endo* zum $C_{2'}$ *endo* wechselt. Solche pucker-Zustandsänderungen erfolgen auch bei den Übergängen von helikalen Segmenten zu ungepaarten loop-Regionen.

Eine charakteristische Abweichung von der helikalen Konformation entsteht ferner wenn die α (P-$O_{5'}$) und η ($O_{3'}$-P) Torsionswinkel in *ap* (anti periplanar, d.h. 150-180 Grad, d.h. auf gut deutsch: trans) respektive -*sc* (minus syn-clinal, d.h., 270-330 Grad, oder -gauche; normal für η) Bereiche gelangen (siehe Abbildung). Geometrisch bewirkt diese Konfiguration eine scharfe "Kurve" - ein sogenannter π-turn. Die beiden Ribosen, die durch einen derartig konfigurierten Phopshatdiester verbunden sind, zeigen dabei in antiparallele Richtung. (Stabilisiert werden diese Konformationen noch zusätzlich durch das stacken eines dem turn folgenden Phosphats mit einer Base.) Es gibt drei geometrisch unterschiedliche turns mit weiterer Abstufung in ihrer Schärfe. Wir wollen hier nicht weiter in ihre Klassifikation eindringen... Wichtig ist aber dass ein solcher π-turn im Anticodon loop auftritt und zwar am $3'$ Phosphat eines invarianten Nukleotids. Dies scheint eine strukturelle Invariante vieler tRNAs zu sein.

Von besonderem Interesse ist das Anticodon-Triplett. Die drei Anticodon Nukleotide sind in einer Konformation die einer einzelsträngigen RNA Helix entspricht. Man vermutet, dass auf diese Weise bei der Paarung mit dem entsprechenden Codon-Triplett einer mRNA eine Doppelhelix Konformation entsteht.

tRNA hat offensichtlich eine faszinierende und komplizierte Struktur. Sie hält sich allerdings nicht ganz von "alleine" aufrecht. Wie alle Nukleinsäuren ist

tRNA ein Polyanion. Die Gegenwart bestimmter Kationen ist erforderlich, damit tRNA die native Konformation beibehält. Prominentestes Kation ist in diesem Zusammenhang Mg^{2+}. Es bindet sehr stark und kooperativ. Die Anzahl von Anlagerungsstellen variiert von tRNA zu tRNA. In Hefe $tRNA^{Phe}$ befinden sie sich in den nichthelikalen loop Regionen.

5 RNA Struktur und katalytische Aktivität

Bevor wir uns abstrakteren Dingen zuwenden, verweilen wir noch etwas bei der deskriptiven Behandlung von RNA Strukturmotifs und setzen diese in Zusammenhang mit katalytischen und chemischen Wechselwirkungen. Die Motivation dafür ist nicht zuletzt, dass wir die evolutionäre Perspektive auf der molekularen Ebene nicht vergessen sollten. Nach der Entdeckung, dass bestimmte RNAs in Strukturen falten können die katalytisch aktiv sind, hält die Spekulation darüber an, dass RNA Katalyse eine entscheidende Rolle in der Entstehung früher Selbst-Replikation und sequenz-gesteuerter Protein-Synthese gespielt haben muss.

Katalytisch aktive RNAs nennt man auch "Ribozyme" (nach Kruger 1982). Die Fragen sind: Wie bilden diese RNAs ein aktives Zentrum aus? Um welche chemischen Reaktionen dreht es sich und wie werden sie katalysiert?

Da gibt es zunächst "grosse" Ribozyme wie group I und group II Introns und die katalytische RNA Untereinheit der RNase P, sowie "kleine" Ribozyme wie "hammerheads". Es stellt sich die Frage ob eine Mindestlänge erforderlich ist, damit eine RNA eine stabile katalytisch aktive Struktur *in Abwesenheit* des Substrats aufrechterhalten kann. Grosse Ribozyme könnten unabhängig vom Substrat eine "fixe" Struktur haben, während kleine Ribozyme ihre aktive Form möglicherweise nur nach Bindung mit dem Substrat erhalten. Die Frage ist offen.

Die Unterscheidung nach Grösse scheint auch im Hinblick auf den chemischen Mechanismus Sinn zu machen: die drei grossen Ribozyme erzeugen bei der Spaltung von RNA $3'$ Hydroxyl-Enden, während die kleinen $2'$,$3'$ zyklische Phosphat-Enden in den Produkten zurücklassen.

Group I introns enthalten einige Abschnitte invarianter Nukleotide. Alle bekannten group I Sequenzen lassen sich in Sekundärstrukturen falten, die eine gemeinsame "core" Region besitzen. An dieser sind die helikalen Bereiche

P3, P4, P6 und P7 (siehe Abbildung) beteiligt. Man beachte den helikalen Bereich aus Pseudoknoten.

In der Spleiss-Reaktion schneidet das Intron sich selbst aus dem RNA Transkript heraus. Die group I Reaktion läuft in zwei Phasen ab: (1) nukleophiler Angriff der $3'$ OH Gruppe eines exogenen Guanosins auf den $5'$-seitigen Spleiss-Punkt. Damit wird die Kette geschnitten. Man beachte, dass nach diesem Schritt das exogene G nun am $5'$ Ende des Introns hängt. (2) Der $5'$ seitige Spleiss-Punkt wird mit dem $3'$ seitigen Spleiss-Punkt ligiert und setzt dabei die Intron-Sequenz frei. Die $5'$ seitige Spleiss-Stelle befindet sich in der gepaarten Region P1, d.h. ein kurzes Segment des $5'$ Exons ist bis zur Spleiss-Stelle mit dem katalytisch aktiven Intron gepaart. Genau dieser Sachverhalt verleiht dem Intron Spezifizität. Manchmal enthält diese Region am Intron auch einen kurzen Abschnitt der zur $3'$ Spleiss-Stelle komplementär ist. Damit können die zu ligierenden Enden zusammengehalten werden. Manche group I Introns benötigen Proteine um ihre Spleiss-Aktivität zu entfalten. Diese Proteine können sogar im Intron selbst codiert sein. In allen bekannten Fällen ist aber der Reaktionsablauf der gleiche.

Group I Introns wurden in verschiedenen Eukaryoten gefunden (precursor Transkripte für mRNA, rRNA, tRNA), aber auch in Eubakterien. In manchen *Tetrahymena* Spezies befinden sich group I Introns in den Genen der rRNA der grossen ribosomalen Untereinheit. Es ist bemerkenswert, dass in diesen Spezies der phylogenetische Baum des group I Introns verschieden ist von den Verwandtschaftsverhältnissen, die man aufgrund der rRNA Vergleiche erhält. Die Verbreitung von group I Introns ist möglicherweise nicht nur auf konventionelle Vererbungskanäle beschränkt.

Group II Introns sind ebenfalls durch eine gemeinsame Struktur definiert. Es scheinen allerdings weniger Nukleotide konserviert zu sein als im group I Fall. Viele group II Introns können sich *in vitro* selbst spleissen, allerdings nur unter unphysiologischen Bedingungen. Das zeigt aber immerhin dass die Reaktion allein von der RNA durchgeführt werden kann. Um die Spleiss-reaktion *in vivo* zu erhalten sind Proteinkomponenten nötig. Der Spleissmechanismus ist den group I Introns im Prinzip ähnlich. Auffallendster Unterschied ist das Nukleophil: es handelt sich um ein *endogenes* Adenin ist. Daraus folgt, dass das Intron beim nukleophilen Angriff zyklisiert wird; man nennt es dann ein "lariat" Intron. Alle bekannten group II Introns wurden in eukaryotischen Organellen gefunden.

Von den grossen katalytischen RNAs gleicht die M1-Untereinheit der RNase P am ehesten einem Enzym - im Sinne, dass es "umsetzt", oder einen "turn over" erzeugt. Die Reaktion, die durch RNase P katalysiert wird, besteht in der Entfernung einer Sequenz vom 5′ Ende eines tRNA precursors. RNase P spaltet durch Hydrolyse und erzeugt 5′ Phosphat Enden und 3′ OH Enden. Das ist dieselbe chemische Spezifizität der group I und II Introns; die meisten Ribonukleasen spalten mit umgekehrter Spezifizität. Das Holoenzym besteht aus einer 377 Nukleotide langen RNA und einem 14 kD Polypeptid. Das Protein ist für *in vivo* Funktionalität erforderlich. Die RNA-Untereinheit reicht aber zur Katalyse unter hohen Salzkonzentrationen aus. Das Polypeptid scheint die Reaktionsbedingungen auf physiologische Werte einzustellen. Weiters scheint das Protein den geschwindigkeitsbestimmenden Schritt bei Mehrfachumsetzungen zu verändern. Der geschwindigkeitsbestimmende Schritt bei der RNA-only Reaktion ist die Freigabe des Produkts.

Es ist klar, dass katalytische Aktivität - "Funktion" - mit der räumlichen Form einer RNA zusammenhängt. Was sich allerdings immer mehr herausstellt ist, dass die Regeln oder Prinzipien nach denen eine aktive Form "gebaut" wird sich zwischen Proteinen und RNAs stark unterscheiden. An der Sekundärstruktur von Proteinen sind Wassertsoffbrücken zwischen Abschnitten des Peptidrückgrates beteiligt. Das führt unter anderem dazu, dass die Aminosäure-Reste nach "aussen" gewandt sind - realtiv zu den "backbone-backbone" Wechselwirkungen. Die Sekundärstruktur der Nukleinsäuren kommt durch Wasserstoffbrückenbindung zwischen den Basen zustande und bringt dadurch ein repetitives Phosphatrückgrat an die Aussenseite. In einem gewissen Sinn ist die Sekundärstruktur der Nukleinsäuren im Vergelich zu jener der Proteine "umgestülpt": was bei den einen "Innen" ist, ist bei den anderen "Aussen". Das hat Konsequenzen im Hinblick auf die Erzeugung von Strukturen höherer Ordnung. Die Tertiärstruktur von Proteinen entsteht durch das Zusammen-Packen von Aminosäure-Seitenketten. Dieser Vorgang wird vornehmlich durch hydrophobe Wechselwirkungen stabilisiert und auch durch Wasserstoffbrücken unterstützt. Um die Tertiärstruktur von RNAs zu "bauen", müssen gepaarte Regionen - helikale Elemente - auf *spezifische* Weise zusammengebracht werden. Das erscheint aber problematisch. Zum Einen sind die Helices Polyanionen; daher kann der Packungsvorgang nicht durch hydrophobe Wechselwirkungen angetrieben werden und muss zudem elektrostatische Abstossung überwinden. Zum Anderen sind Phosphatrückgrate chemisch gesehen "arm" im Vergleich zu den Basen. Letztere sind aber in der

Helix versteckt und nur teilweise tertiären Wechselwirkungen zugänglich. Die Lösung des Problems besteht wie bei den Proteinen teilweise aus Wasserstoffbrücken über das Rückgrat, Pseudoknoten, Basentripel und dergleichen wie wir sie im Fall der tRNA kennengelernt haben. Es gibt aber einen weiteren wichtigen Beitrag zur Lösung des Tertiärstrukturproblems in RNA. Wir sind ihm flüchtig bei der tRNA schon begegnet: Metallkationen.

Es ist der Beitrag der Metallkationen, der die Protein-3D-Faltung von der RNA-3D-Faltung unterscheidet. Die meisten Proteine falten ohne Metallionen; mit Ausnahmen wie etwa die "Zink-Finger". tRNA, das *Tetrahymena* Ribozym (group I Intron) und viele weitere benötigen Mg^{2+} oder Mn^{2+}. Die Kationen wirken wahrscheinlich auf der Ebene der Tertiärtstruktur eher als auf der Ebene der Sekundärstruktur. Bei der tRNA scheinen die divalenten Kationen Regionen zu stabilsieren in denen sich zwei oder mehr Phosphatrückgrate nahe kommen.

Ein weiterer Unterschied zwischen RNA- und Protein-Strukturprinzipien ist die Autonomie der einzelnen Sekundärstrukturelemente. Wenn eine α-Helix aus ihrem Kontext im Protein herausgenommen wird, "entfaltet" sie sich. Im Gegensatz dazu behält eine RNA Haarnadel ihre Struktur häufig bei. Eine Konsequenz davon ist dass man diskrete Bestandteile der RNA Funktion individuellen Sekundärstrukturelementen zuschreiben kann. Umgekehrt kann man vielleicht eine RNA-Maschine gezielt aus Sekundärstrukturelementen zusammenbauen, von denen jedes einen definierten Beitrag zur Katalyse leistet. Wenn beispielsweise ein loop zur Metallionenbindung mit einem loop zur Guanosinbindung richtig zusammengebracht wird, könnte dies zur Ausrichtung und Aktivierung eines Guanosinnukleophils dienen. Obwohl dieses "Kassettenprinzip" der Tertiärstruktur von RNAs naiv ist, könnte es einen nützlichen Ausgangspunkt zur Betrachtung von Struktur-Funktionsbeziehungen in RNA liefern.

Das Tertiärstruktur-Modell vom *Tetrahymena* Ribozym ist in der beigelegten Abbildung dargestellt und soll hier nicht weiter ausgeführt werden.

Eine weitere Abbildung stellt die bisher bekannten Reaktionen dar, die von Ribozymen katalysiert werden. In der Natur wurden bisher nur Transesterifizierungen und Hydrolyse von Phosphodiester Bindungen gefunden. Der Spaltungsschritt verläuft unter Katalyse des *Tetrahymena* Introns schneller als die Helixbildung bei Substratanlagerung. In diesem Sinne ist die Reaktion "diffusionskontrolliert". Das Ribozym saturiert aber rasch, weil es das

gespaltene $5'$ Ende nur sehr langsam loslässt. Dafür wurde es allerdings auch selektiert...

Die grossen Ribozyme sind in der Lage auch DNA zu spalten; allerdings verläuft die Reaktion um einige Grössenordnungen langsamer. Dass sie überhaupt möglich ist, ist einzusehen, da der nukleophile Angriff bei den grossen Ribozymen nie das $2'$ OH direkt miteinschliesst. Trotzdem beeinflusst die Hydroxylgruppe offensichtlich die Reaktivität auf indirekte Weise stark.

Die kleinen Ribozyme, wie "hammerhead" und hepatitis δ - eine Satelliten RNA des Hepatitis B virus -, stellen RNA Struktur-Motive dar, die RNA Spaltungsreaktionen katalysieren. Sie treten in RNAs von bestimmten pathogenen Pflanzen Viren auf. Die Spaltung hinterlässt, wie bereits erwähnt, ein $2'$,$3'$ zyklisiertes Phosphat Ende. Wiederum sind divalente Metallkationen wesentlich an der Funktion beteiligt. Es soll nicht der Eindruck entstehen, dass RNA Katalyse im Gegensatz zu proteinvermittelter Katalyse "schlampig" sei. Ebenso wie bei tRNA und den grossen Ribozymen verlangt die Aktivität der kleinen Ribozyme die Erfüllung präziser geometrischer Anforderungen.

RNA Strukturelemente und Funktion

Abschliessend fassen wir kurz die Strukturelemente der RNA und einige ihrer Funktionen zusammen.

doppelsträngige helikale Abschnitte

• Sie bewirken thermodynamisch die Strukturbildung, und sind indirekt Ursache für die Entstehung aller anderen Strukturelemente. • Stabilisieren Tertiärstruktur (rRNA, RNase P RNA, group I self-splicing introns). • Spezifizität in der Katalyse. Erkennung eines spezifischen RNA Substrats seitens einer katalytischen oder strukturellen RNA erfolgt weitgehend über komplementäre Helix-Bildung (Spliceosomen, Translation von mRNA - Shine/Dalgarno, anticodon Erkennung, group I introns, antisense Regulation) • Wechselkwirkung mit bestimmten Proteinen: vor allem solchen, die helikale Abschnitte schmelzen.

hairpin loops

• Besonders stabile "Tetraloops", bestehend aus spezifischen Tetraoligonu-

kleotiden. Die zwei mittleren Nukleotide haben $C_{2'}$ *endo* pucker. • Zwei Klassen: solche mit definierter rigider Struktur, und solche mit flexibler Struktur. • Beteiligt an der Initiation, Propagation, und Termination der Translation. Attenuations-Kontrolle. • Kontrolle der mRNA Halbwertszeit. • Pausieren der Ribosome. Frameshifting. • Kontakte mit Proteinen. • Spezifische Erkennung - fällt auch in die Kategorie "doppelsträngige Helices", denn die Bildung gepaarter Regionen impliziert die Bildung ungepaarter (loop-)Regionen.

internal loops

• Kontakte mit Proteinen. Bindung ribosomaler Proteine.

bulges (einseitige Ausbuchtungen)

• Knicke in der Tertiärstruktur. • Bereitstellung eines Nukleophils in katalytischer Aktivität (herausgebuchtetes Adenosin in group II introns).

Pseudoknoten

• Haben A-Form Geometrie. • tRNA Mimicry bei der RNA von Pflanzen-Viren. • Frameshifting bei der Translation retroviraler mRNAs. • Protein-Bindungsstellen.

6 Die Kombinatorik der Sekundärstruktur

Die Primärstruktur einer einzelsträngigen RNA ist eine lineare Sequenz $\mathbf{a} = a_1 a_2 \dots a_n$, mit $a_i \in \{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$. Formal gesehen ist die Sekundärstruktur ein Graph, dessen Knoten die Basen sind und dessen Kanten die Kontakte zwischen den Basen ausdrücken. Dabei gibt es zwei Arten von Kontakten: (1) die Phosphat-Verbindungen zwischen den Basen, die das Rückgrat ausmachen, und (2) die Menge der Basenpaare. Menge (1) ist uninteressant, weil sie ohnehin durch die Sequenz festgelegt ist. Es ist die Menge der Basenpaare, Menge (2), die eine Sekundärstruktur charakterisiert. Eine Sequenz kann nämlich viele mögliche Basenpaarungen (Sekundärstrukturen) einnehmen. Diese werden sich zwar energetisch unterscheiden, aber damit wollen wir uns noch nicht befassen.

Definition 6.1. *Eine Sekundärstruktur besteht aus einer Knotenmenge* $V = \{1, 2, \dots, i, \dots, N\}$ *und einer Menge* S *von Kanten* $i \cdot j$, $1 \leq i < j \leq N$, *derart, dass*

(i) (Rückgrat) $\forall\, i < n\; i \cdot (i+1) \in S$*, und*
(ii) (binäre Paarungen) für jedes i *gibt es höchstens ein* $k \neq i-1, i+1$ *für das* $i \cdot k \in S$*, und*
(iii) (Basenpaare ohne Pseudoknoten) falls $i \cdot j \in S$ *und* $k \cdot l \in S$ *und* $i < k < j$*, dann gilt* $i < l < j$*.*

Die Knotenmenge steht einfach für die Menge der Positionen einer Sequenz der Länge N. Man beachte, dass die Kantenmenge das Rückgrat der Kette, d.h. Kontakte zwischen entlang der Kette benachbarten Basen, miteinschliesst. Die Kantenmenge besteht also aus Rückgrat + Basenpaare. Wir wollen aber in Hinkunft das triviale Rückgrat vergessen und so tun als bezeichne $i \cdot j$ ein Basenpaar. Dabei ist es nützlich ein bürokratisches Detail zu bemerken: die obige Definition schliesst Basenpaarung zwischen aufeinanderfolgenden Positionen i und $i+1$ aus. Das ist deshalb der Fall, weil Positionen i und $i+1$ immer durch eine Rückgrat-Kante verbunden sind und die Definition keinen Unterschied zwischen Rückgrat-Kante und Basenpaar-Kante macht. Das ist aber völlig egal, weil es ohnehin sterisch nicht sinnvoll ist i und $i+1$ miteinander paaren zu lassen. Wenn man dies jedoch zulässt und dazu noch das triviale Rückgrat vergisst, könnte die Definition etwas einfacher aussehen: Eine Sekundärstruktur ist eine Menge von Basenpaaren (ohne Pseudoknoten): $S' = \{i \cdot j | \forall i \cdot j \quad \nexists\, k \cdot l \quad i \leq k \leq j \leq l\}$

Die Definition fällt deshalb etwas schwerfällig aus, weil sie Pseudoknoten ausschliesst. Der Ausschluss von Pseudoknoten entspricht dem üblichen Begriff der Sekundärstruktur. Dieser Begriff ist durchaus sinnvoll - wie wir bereits in den vorangegangenen Abschnitten gesehen haben. Man beachte ferner, dass die Sekundärstruktur ein topologischer Begriff ist. Er bringt zum Ausdruck welche Position welcher benachbart ist. Es wird gar nichts darüber ausgesagt wie weit eine Position von einer anderen entfernt ist. Dies sollte man stets gegenwärtig halten, insbesondere bei Betrachtung graphischer Darstellungen der Sekundärstruktur, die einen verleiten können anzunehmen es handle sich um die "2D-Darstellung" einer Struktur. Das ist nicht der Fall!

In der formalen Definition 6.1 einer Sekundärstruktur wird offengelassen was man unter "Basenpaar" versteht. Die Definition macht selbst dann Sinn wenn jede Base mit jeder paaren kann. Üblicherweise fügt man eine Liste zugelassener Paarungen hinzu.

Bleiben wir zunächst bei der uneingeschränkten Definition 6.1, d.h. ignorieren wir die spezifische Basensequenz einer RNA Kette und lassen Paarungen

zwischen beliebigen Nukleotidsorten zu. Die Frage, die sich dabei aufdrängt, lautet: Wie viele unterschiedliche Sekundärstrukturen - d.h. unterschiedliche Topologien - gibt es für eine Kette der Länge N? Das wollen wir uns näher ansehen, weil man damit ein Gefühl für den "rekursiven" Aufbau einer Struktur erhält.

Sei die gesuchte Anzahl an Sekundärstrukturen der Länge n, $S(n)$. Wenn $n = 0$, erlauben wir uns einfach zu definieren: $S(0) = 1$. Für $n = 1$ und $n = 2$ ist sofort klar, dass $S(1) = 1$. Ein direktes Abzählen wird bald sehr frustrierend. Versuchen wir also ein Verfahren zum Abzählen zu entwickeln. Nehmen wir an, wir kennen die Zahl der Strukturen für alle Längen bis zur Länge l, $S(k), 1 \leq k \leq l$. Verlängern wir nun die Kette $[1, l]$ um eine Position zur Kette $[1, l + 1]$. Wir können dann die Zahl der Strukturen auf $[1, l + 1]$ durch die bekannte Zahl von Strukturen auf Teilsegmenten ausdrücken.

Es gibt zwei Möglichkeiten für das neu hinzugekommene Nukleotid $l + 1$: Entweder $l + 1$ ist nicht gepaart, oder es paart mit einer Position j, wobei $1 \leq j \leq l - 1$.

Wenn $l + 1$ nicht gepaart ist, dann gibt es $S(l)$ Strukturen, die das Segment $[1, l]$ annehmen kann.

Nehmen wir nun an, $l+1$ paare mit Position j. Durch diese Paarung entstehen zwei Teilsegmente $[1, j-1]$ (Länge $j-1$) und $[j+1, l]$ (Länge $l-j$), die beide ihren eigenen unabhängigen Satz an Strukturen einnehmen können, nämlich $S(j-1)$ und $S(l-j)$. Jede Struktur auf einem Segment ist mit jeder Struktur auf dem anderen frei kombinierbar und ergibt eine mögliche Struktur auf der Gesamtkette, insgesamt also $S(j-1)S(l-j)$. Das gilt aber für jedes mögliche j, $1 \leq j \leq l - 2$.

Was wir gerade hergeleitet haben, können wir gleich als Theorem zusammenfassen:

Theorem 6.1. *Für die Zahl $S(l+1)$ der Sekundärstrukturen einer Kette der Länge $l + 1$ gilt:*

$$S(l + 1) = S(l) + \sum_{j=1}^{l-1} S(j - 1)S(l - j)$$

Auf diese Weise kann man rekursiv die Zahl der Sekundärstrukturen bestimmen: $1, 1, 2, 4, 8, 17, 37, 82, 185, \ldots$. Für eine Kette der tRNA Länge $N = 76$

müssten wir sukzessive $S(1)$, $S(2)$, usw. bis $S(75)$ ausrechnen um dann endlich $S(76)$ zu bestimmen.

Wenn wir die Definition S' benutzen, in der Basenpaarung zwischen i und $i+1$ zugelassen ist, dann können wir mit (fast) elementarer Kombinatorik eine exakte Abzählung durchführen. Das wollen wir des pädagogischen Wertes wegen auch tun.

Verglichen mit $S(l+1)$ ändert sich nur die Summationsobergrenze: $S'(l+1) = S'(l) + \sum_{j=1}^{l} S'(j-1)S'(l-j)$. Die Beziehung zwischen $S(l+1)$ und $S'(l+1)$ ist wegen der Rekursion dennoch nicht einfach.
(1) Zunächst stellen wir fest dass $S'(n) = \sum_{k=0}^{\lfloor n/2 \rfloor} S'(n,k)$ ist, wobei $S'(n,k)$ die Zahl der Sekundärstrukturen von Ketten der Länge n mit genau k Basenpaaren ist.
(2) $S'(n,k)$ lässt sich auf folgende Weise berechnen. Wir stellen eine Sekundärstruktur als eine Folge von Punkten (ungepaarte Positionen) und gepaarten Klammern (gepaarten Positionen) dar, beispielsweise ...((..))..(((..(.))..))
Eine Sekundärstruktur mit k Basenpaaren hat natürlcih k Klammernpaare. Wir vergessen vorläufig die $n-2k$ Punkte (n ist die Kettenlänge) und fragen wieviele Anordnungen von k Klammernpaaren gibt es? Die Frage ist identisch mit der Frage nach den möglichen vollständigen Klammerungen eines Produkts mit $n+1$ Termen, $a_0 \cdot a_1 \cdot \ldots \cdot a_n$. Die Antwort sind die Catalan Zahlen:

$$\frac{1}{k+1}\binom{2k}{k} \qquad (\sim 4^k \cdot k^{-3/2})$$

Jetzt müssen wir noch die $n-2k$ Punkte unterbringen. Betrachten wir eine Klammeranordnung, etwa ((()))(), dann können wir Punkte vor jeder aufgehenden und vor jeder zugehenden Klammer einsetzen, sowie am Ende der Klammeranordnung. Insgesamt sind das $2k+1$ Plätze. Wir wissen (oder können leicht ableiten) dass die Zahl der Möglichkeiten a (ununterscheidbare) Objekte auf b Schachteln zu verteilen $\binom{a+b-1}{a}$ ist (Bose-Einstein!). Im vorliegenden Fall müssen wir $n-2k$ Punkte auf $2k+1$ Plätze verteilen. Das gibt

$$\binom{n-2k+2k+1-1}{n-2k} = \binom{n}{n-2k}$$

Möglichkeiten pro Klammeranordnung. Insgesamt also:

$$S'(n) = \sum_{k=0}^{\lfloor n/2 \rfloor} \frac{1}{k+1}\binom{2k}{k}\binom{n}{n-2k}$$

Kehren wir zu Theorem 6.1 zurück. Für diese Rekursion lässt sich eine asymptotische Formel angeben. Die Herleitung der Asymptotik erfolgt über erzeugende Funktionen, $\phi(x) = \sum_{k\geq 0} S(k)x^k$, und benötigt einige Tricks, die hier zu erläutern uns zu lange beschäftigen würde. (Daher rührt der pädagogische Wert der vorigen Abzählung :-) Das Egebnis lautet:

Theorem 6.2. *Für $n \to \infty$ ist die Zahl $S(n)$ der Sekundärstrukturen:*

$$S(n) \sim \sqrt{\frac{15+7\sqrt{5}}{8\pi}}\, n^{-3/2} \left(\frac{3+\sqrt{5}}{2}\right)^n \approx 1.1\, n^{-3/2}\, 2.6^n$$

Für eine bescheidene 5S RNA mit 120 Nukleotiden sind das etwa $1.2 \cdot 10^{47}$ verschiedene Sekundärstrukturen. Wir wollen uns aber nicht von den Zahlen ablenken lassen. Wir stellen fest, dass die Zahl der Strukturen exponentiell in der Kettenlänge ist. Nun, das war irgendwie zu erwarten. Das Interessante aber an der asymptotischen Formel ist der Term 2.6^n im Vergleich zu 4^n - die Zahl der Sequenzen der Länge n mit 4 Buchstaben. Die Asymptotik sagt uns, dass es viel weniger Strukturen gibt als Sequenzen! Man vergesse nicht, dass Theorem 6.2 die Strukturen zählt *ohne Einschränkung durch Basenpaarregeln*. Jede Position kann mit jeder paaren (ausgenommen benachbarte, s.o.). Die Zahl ist also eine (schlechte) obere Schranke zu dem was tatsächlich biophysikalisch realisiert werden kann. Wenn man davon ausgeht, dass nur Energieminima eingenommen werden, dann wird diese Zahl noch weiter sinken.

Ein stochastischer Ansatz bei dem wir die Watson-Crick Regeln berücksichtigen liefert eine interessante Verfeinerung des obigen Resultats. Wir erzeugen eine RNA-Sequenz der Länge n mit den Nukleotid-Wahrscheinlichkeiten p_A, p_U, p_G, p_C. Die Wahrscheinlichkeit, dass zwei beliebig herausgegriffene Positionen miteinander paaren können ist $p = 2(p_A p_U + p_G p_C)$ Sei $\eta_{ij} = 1$, wenn i mit j paaren kann, 0 sonst. Offensichtlich ist der Erwartungswert $\langle \eta_{ij} \rangle = p$. Sei nun $R(n)$ eine Zufallsvariable, die die Zahl der Sekundärstrukturen einer Zufallssequenz der Länge n bezeichnet. Wir erhalten ganz analog zur vorangehenden Überlegung:

$$R(n+1) = R(n) + \sum_{j=1}^{n-1} R(j-1)R(n-j)\eta_{l+1,j}$$

Wobei die Multiplikation mit $\eta_{l+1,j}$ dafür sorgt, dass nur jene Produkte von Teilsegmenten eingehen für die Position $l+1$ tatsächlich mit Position j paaren

kann. Wir bilden nun den Erwartungswert (man beachte, dass die Teilsegmente unabhängig sind)

$$\langle R(n+1)\rangle = \langle R(n)\rangle + \sum_{j=1}^{n-1} \langle R(j-1)\rangle \langle R(n-j)\rangle p$$

Für diese Rekurrenz ist eine analoge asymptotische Berechnung möglich. Das Resultat liefert dabei eine ähnliche exponentielle Abhängigkeit von n wie vorhin. Im Fall dass alle Nukleotide gleichverteilt sind, $p = 1/4$, lautet der exponetielle Term: 1.86^n. Diese Abschätzung ist interessant, weil sie zeigt, dass die (mittlere) Zahl der Sekundärstrukturen für Sequenzen der Länge n sogar kleiner ist als die Zahl 2^n der Sequenzen über dem kleinst-möglichen Alphabet, einem binären Alphabet.

Jede RNA Sequenz der Länge n muss also irgendeine der höchstens 2.6^n (oder stochastisch 1.86^n) möglichen Sekundärstrukturen annehmen. Das heisst: viele Sequenzen werden in diesselbe Sekundärstruktur falten. Man beachte, dass diese Sequenz/Struktur-Entartung anders ausfallen wird, je nach dem wie man "Struktur" definiert. Wenn wir unter "Struktur" die exakten 3D-Koordinaten jedes Atoms in einer RNA verstehen, dann haben keine zwei Sequenzen dieselbe Struktur. Eine solcherart definierte Struktur ist natürlich sehr schwierig (wenn überhaupt) zu berechnen. Er ist aber andererseits für eine Reihe von Fragen (auf die wir später noch kurz zu sprechen kommen) gar nicht nötig. Wie sich aus den Beispielen der Vortage zeigt, kann man biologische Funktionalität bereits mit einem wesentlich weniger detailliertem, extrem groben Strukturbegriff – wie dem der Sekundärstruktur – sinnvoll korrelieren. Halten wir also fest, dass es nicht nur einen einzigen nützlichen Strukturbegriff gibt, sondern eine Reihe sinnvoller Abstraktionen. "Sinnvoll" heisst hier eine Mischung aus "hat Erklärungswert", "ist empirisch zugänglich", "kann berechnet werden".

Es ist klar, dass nicht alle Sekundärstrukturen dieselbe freie Bildungsenergie aufweisen werden. Einige Strukturen werden stabiler sein als andere. Wegen der stacking Wechselwirkungen wird beispielsweise eine vollständig gepaarte Haarnadel stabiler sein als eine Struktur mit einem einzigen isolierten Basenpaar. Wir wenden uns nun der Berechnung einer energetisch optimalen Sekundärstruktur einer beliebigen RNA Sequenz zu.

Zur Vorbereitung soll eine kurze Betrachtung eingeschaltet werden, die das Prinzip nach dem vorgegangen wird, verdeutlicht. Statt die Sekundärstruk-

tur mit niedrigster Energie zu suchen, wollen wir ein Verfahren angeben um zu einer gegebenen Sequenz eine Sekundärstruktur mit der maximal möglichen Anzahl von Basenpaaren zu finden. Zu diesem Zweck brauchen wir eine Matrix, die uns angibt welche Basenpaare erlaubt sind, z.B. $A \cdot U$, $C \cdot G$, $U \cdot A$ usw. Sei das Alphabet $\mathcal{A} = \{A, U, G, C\}$ und $a, b \in \mathcal{A}$, dann wollen wir:

$$\rho(a,b) = \begin{cases} 1 & \text{wenn } a \text{ und } b \text{ paaren können,} \\ 0 & \text{sonst.} \end{cases}$$

Wir stellen nun eine Überlegung an, der wir bereits bei der Herleitung der Rekursionsformel für die Anzahl an Sekundärstrukturen begegnet sind. Sei eine Nukleotidsequenz $a_1 a_2 \cdots a_n$ mit $a_i \in \mathcal{A}$ gegeben. Bezeichne ferner $X_{i,j}$ die maximale Zahl an Basenpaaren im Teilsegment $[i,j], i < j$ (also der Teilsequenz $a_i a_{i+1} \cdots a_j$). Nehmen wir weiter an wir kennen $X_{i,j}$ und wollen $X_{i,j+1}$ berechnen, d.h. wir hängen das $j+1$-te Nukleotid an das $3'$ Ende des bereits berechneten Segments und erzeugen das Segment $[i, j+1]$. Wir gehen nun alle möglichen Wechselwirkungen durch, die das Nukleotid a_{j+1} eingehen kann. Die Position $j+1$ kann im Prinzip mit jeder Position l, $i \leq l <\leq j-1$ wechselwirken. Ob sie paaren kann, sagt uns das Matrixelement $\rho(a_{j+1}, a_l)$. Wenn das Nukleotid an der Stelle $j+1$ mit dem an der Stelle l paaren kann, dann wird das Segment $[i, j+1]$ in zwei unabhängige (keine Pseudoknoten!) Teilsegmente $[i, l-1]$ und $[l+1, j]$ gespalten. Wenn $X_{i,l-1}$ und $X_{l+1,j}$ die maximale Anzahl an Basenpaaren auf den entsprechenden Teilsegmenten bezeichnen, dann ist die Gesamtzahl an Basenpaaren auf dem Segment $[i, j+1]$ im Fall dass l und $j+1$ paaren, genau $1 + X_{i,l-1} + X_{l+1,j}$. Wir gehen alle Fälle für $i \leq l \leq j-1$ durch und behalten den Maximalwert. Diesen vergleichen wir schliesslich mit $X_{i,j}$ und der grössere der beiden ist der Wert für $X_{i,j+1}$. Der letzte Vergleich ist nötig, weil es ja sein könnte, dass eine Paarung von $j+1$ wegen der Aufspaltung in Segmente uns im Endeffekt Basenpaare kostet - jene nämlich die vorher zwischen diesen Segmenten möglich waren; es könnte also besser sein $j+1$ gar nicht paaren zu lassen. Wir fassen nun diesen Verbalexzess zusammen:

$$X_{i,j+1} = \max\{X_{i,j}, \max_{i \leq l \leq j-1} \{[X_{i,l-1} + 1 + X_{l+1,j}]\rho(a_l, a_j + 1)\}\}$$

Wir müssen aber noch eine Annahme einlösen, die wir immer wieder gemacht haben. Wir haben so getan als wären die $X_{p,q}$ der Teilsegmente schon berechnet worden bevor wir sie brauchten. Genauer: wenn wir $X_{i,j}$ berechnen, brauchen wir in der $(i \leq l \leq j-1)$-Schleife alle Eintragungen $X_{p,q}$ mit

$i \leq p < q \leq j-1$. Der Witz ist, dass man die $n \times n$ Matrix der $X_{p,q}$ Werte so auffüllen kann, dass wir nie auf Eintragungen zugreifen, die noch nicht berechnet worden wären. (Es handelt sich übrigens nur um eine Hälfte der Matrix, da nur jene $X_{p,q}$ einen Sinn haben für die $p < q$ ist). Beispielsweise kann man von der Hauptdiagonalen her alle Nebendiagonalen auffüllen, d.h. zunächst werden alle Segmente $[i, i+2], i \leq 1 \leq n-2$ berechnet, dann alle Segmente $[i, i+3], i \leq 1 \leq n-3$, usw. bis zum letzten - dem gesuchten - Segment $[1, n]$.

Wenn die Prozedur beendet ist, haben wir eine Zahl: die maximal mögliche Zahl m an Basenpaarungen. Wir hätten aber gern auch eine Liste dieser Paare, d.h. eine Sekundärstruktur mit m Basenpaaren. Das ist leicht, denn wir haben nicht nur eine Zahl ausgerechnet, sondern auch eine ganze Tabelle von $X_{i,j}$. Wir drehen jetzt einfach den Algorithmus um: Wir starten mit dem Segment $[1, n]$ (der gesamten Kette) und schauen in der Tabelle nach wie wir auf diese Zahl gekommen sind, d.h. welche Segmentierung in die Teilketten $[1, l-1]$ und $[l+1, n-1]$ den Wert $X_{1,n}$ geliefert hat. Wenn eine Segmentierung gefunden wurde, dann war dafür das Paar $l \cdot k$ verantwortlich und wir nehmen es in die Liste auf. Genauso verfahren wir mit den Teilsegmenten $[1, l-1]$ und $[l+1, n-1]$ bis nichts mehr übrigbleibt. Die so erhaltene Liste ist nicht notwendigerweise die Einzige mit der maximalen Zahl an Paarungen. Es wird im allgemeinen viele entartete optimale Lösungen geben. Was wir mit dieser Prozedur getan haben, ist *eine* dieser optimalen Lösungen zu generieren. Dieses Verfahren, eine Tabelle, die durch einen Algorithmus aufgefüllt wurde, durch Umkehr desselben Algorithmus zu durchschreiten nennt man "backtracking". Das geht sehr schnell.

Diese Strategie eine Tabelle aufzufüllen heisst "dynamic programming". Dynamic Programming ist eine ausserordentlich wichtige und vielseitig anwendbare Technik um Optimierungsprobleme zu lösen, indem Lösungen zu Teilproblemen gefunden werden und diese dann zur Gesamtlösung zusammengefügt werden. Das Wort "programming"hat hier *nichts* mit der Erstellung von Computer-Code zu tun, sondern bezeichnet lediglich das Auffüllen einer Tabelle; "programming" heisst hier soviel wie "Tabellierung". Das Prinzip stammt halt aus einer Zeit wo der Unterschied noch nicht so offenkundig war...

Es ist lehrreich die Zahl der Schritte grob abzuschätzen, die dieser Algorithmus benötigt. Für jedes i und jedes j, $i < j$, durchlaufen wir eine Schleife

$j-i-1$ mal (das $\max_{i \leq l \leq j-1}$ in der obigen Gleichung). Das können maximal n Schritte sein. Das tun wir aber für jedes i und jedes j. Das sind wiederum ordnungsmässig n^2 Schritte. Insgesamt also n^3. Das ist beachtlich, wo wir doch vorhin ausgerechnet haben, dass die Grösse unseres Suchraums - das ist die Anzahl möglicher Sekundärstrukturen - der Ordnung c^n - also exponentiell ist! Mit einem naiven Vergleichen aller Sekundärstrukturen wären wir Jahrhunderte und länger unterwegs. Dynamic programming hat also ein exponentielles Problem in ein polynomiales verwandelt. Das ist keine Magie, denn es funktioniert auch nicht bei allen Optimierungsproblemen.

Die Anwendbarkeit von dynamic programming verlangt, dass ein Optimierungsproblem eine ganz bestimmte Struktur aufweist. Ein Optimierungsproblem kann als eine Folge von Entscheidungen gesehen werden. In unserem Fall ging es darum zu entscheiden welche Position mit welcher paart. Eine *Folge* von Entscheidungen heisst eine *optimale Folge* wenn sie folgender rekursiven Definition genügt: Greifen wir einen Zustand in dieser Folge heraus und nennen wir ihn den "gegenwärtigen Zustand". Die Eigenschaft lautet: unabhängig davon welche Entscheidungen in der Vergangenheit zum gegenwärtigen Zustand geführt haben, muss relativ zu diesem Zustand die restliche Folge von Entscheidungen auch eine optimale Folge sein. Das bedeutet im Wesentlichen dass das Problem in Teilprobleme zerlegbar sein muss. Das heisst nicht, dass diese Teilprobleme voneinander unabhängig sein müssen! Sind sie es auch nicht: Die Probleme eine optimale Paarung auf dem Segment $[10, 50]$ und eine optimale Paarung auf dem Segment $[30, 70]$ zu finden, haben etliche Teilprobleme gemeinsam. Zerlegbarkeit bedeutet hier vielmehr, dass eine Entscheidung zum Zeitpunkt t nicht die Optimalität einer vorangegangenen Entscheidung zerstört. Das ist natürlich weder im Leben so, und auch nicht bei der Berechnung von echten Molekül-Konformationen, wie etwa Proteinstrukturen! Wegen der starken Kontextabhängigkeit kann auch dynamic programming nicht helfen.

Im Fall der Basenpaarmaximierung ist die Zerlegbarkeit durch die simple Additivität der Basenpaaranzahlen garantiert. Das nützt der erläuterte Basenpaarmaximierungsalgorithmus auf einfache aber clevere Weise aus. Die Anwendbarkeit von dynamic programming auf RNA Sekundärstrukturen wurde am Ende der 70er Jahre von mehreren gleichzeitig gesehen.

Wir sind ausreichend gerüstet um uns nun kurz der Bestimmung einer energetisch optimalen Sekundärstruktur zuzuwenden. Im Fall der Basenpaarma-

ximierung war das einzelne "Basenpaar" offensichtlich der angemessene elementare Baustein in den Sekundärstrukturen zerlegt werden können. Wenn wir Energie ins Spiel kommt, ist dies nicht mehr der Fall. Der Grund ist ganz einfach. Z.B. hat ein Basenpaar in einem "stack" eine andere Energie als wenn es eine offene Schleife abschliesst. Mit energetischen Betrachtung kommt unvermeidlich eine gewisse Kontextabhängigkeit ins Spiel. Diese kann man bei RNA Sekundärstrukturen in vernünftigen Grenzen halten. Die Vorgangsweise besteht darin einen sinnvollen Satz von Strukturelementen zu definieren, in die jede Sekundärstruktur zerlegt werden kann und zu postulieren, dass jedes Strukturelement additiv zur Energie der Gesamtstruktur beiträgt. Die Additivität ist es, die es wiederum ermöglicht dynamic programming anzuwenden.

Das angemessene Struktureinheit ist eine "Schleife" oder ein **loop**, wie wir sagen werden. Ein loop ist durch zwei Zahlen charakterisiert: die Anzahl der Helices, die von ihm ausgehen - diese Zahl wollen wir den "Grad" des loops nennen - und die Anzahl ungepaarter Nukleotide, die er enthält - seine "Grösse". Dabei ist es sinnvoll Unterscheidungen zu treffen und wir wollen diese Unterscheidungen als die "fundamentalen Strukturelemente" betrachten:

hairpin: ist ein loop vom Grad 1 und beliebiger Grösse

internal loop: ist ein loop vom Grad 2 und beliebiger Grösse (Dabei trifft man häufig eine weitere Unterscheidung zwischen einem echten internal loop, in dem beidseitig ungepaarte Nukleotide vorkommen, und in eine "Ausbuchtung" ("bulge") bei der die ungepaarten Nukleotide nur einseitig anzutreffen sind.)

stack: das sind zwei unmittelbar aufeinanderfolgende Basenpaare, $i \cdot j$ und $(i+1) \cdot (j-1)$. Das ist ein loop vom Grad 2 und der Grösse 0.

multiloop: das ist ein loop mit Grad > 2 und beliebiger Grösse.

externe Elemente: Das sind die ungepaarte Abschnitte, die zu keinem loop gehören. Etwa Segmente ungepaarter Nukleotide, die Teilstrukturen verbinden, sowie freie $3'$ oder $5'$ Enden. Alle externen Elemente werden zu einem Strukturelement (variabler Grösse zusammengefasst). Man könnte sich bei einer Kette $[1, n]$ ein "virtuelles" Basenpaar zwischen

den virtuellen Positionen 0 und $n+1$ vorstellen. Der dadurch entstandene loop umfasst genau die externen Elemente.

Weil Pseudoknoten ausgeschlossen sind und weil jede Base an höchstens einer Basenpaarung teilnehmen kann, kann jede Position (jedes Nukleotid) in einer Sekundärstruktur genau einem dieser Strukturelemente zugeordnet werden.

Wir postulieren nun, dass jedes dieser Elemente additiv einen Energiebeitrag (der von seiner Grösse und möglicherweise anderen Parametern abhängt) zur Gesamtenergie beiträgt. Sei nun $E(S)$ die minimale Energie und S die dazugehörige Struktur. Sei ferner $i \cdot j$ ein Basenpaar von S und sei S_{ij} eine Sekundärstruktur auf dem Teilsegment $[i, j]$. Aus der postulierten Additivität folgt, dass die Teilstruktur S_{ij} ebenfalls optimal sein muss. Wäre nicht S_{ij} optimal, sondern eine andere Struktur auf demselben Segment, sagen wir S'_{ij}, dann könnten wir S_{ij} durch S'_{ij} ersetzen und damit eine Struktur mit geringerer Gesamtenergie als S erhalten. Das widerspricht aber unserer Annahme, dass S optimal ist! Mit anderen Worten: Unter der Voraussetzung der Additivität muss bei einer optimalen Sekundärstruktur jede Teilstruktur, die von einem Basenpaar abgeschlossen wird, ebenfalls optimal sein. Das Optimalitätsprinzip – die Voraussetzung für die Anwendbarkeit von dynamic programming – ist also erfüllt.

Für jedes der oben angeführten Strukturelemente gibt es empirische Messwerte zur freien Energie. Diese Energien sind tabelliert und zwar in Abhängigkeit der Grösse der Schleife, der Zahl der Helices, die in sie münden, und der Art der Basenpaare die eine Schleife begrenzen. Solche Messungen sind nicht einfach durchzuführen und sie sind auch nicht immer mit einer hohen Genauigkeit behaftet. Jedenfalls sind diese Tabellen ein äusserst nützlicher Datengrundstock, der Sekundärtsrukturberechnungen eine gewisse Verankerung in der empririschen Wirklichkeit verleiht. Die stacking Energien sind dabei die einzigen stabilisierenden Energiebeiträge, Schleifen sind destabilisierend.

Die Prozedur zur Energieminimierung verläuft nun ganz analog wie wir sie von der Basenpaarungsmaximierung her kennen. Der einzige wesentliche Unterschied ist dass wir es nun mit einem grösseren Satz an elementaren Strukturbausteinen zu tun haben. Sei $e(s)$ die Energie eines elementaren Strukturbausteins s und sei $E_{i,j}$ die minimale Energie auf dem Abschnitt $[i, j]$ wenn

i mit j paart. Dann ist

$$E_{i,j} = \min_{\substack{\text{alle in Frage} \\ \text{kommenden } s}} \left\{ \min_{\substack{s \text{ ist ein Struk-} \\ \text{turelement, das} \\ \text{durch } i \cdot j \text{ abge-} \\ \text{schlossen wird}}} \{e(s) + \sum_{\substack{\text{alle } p \cdot q \text{ aus-} \\ \text{ser } i \cdot j, \text{ die} \\ \text{das Struktur-} \\ \text{element } s \text{ be-} \\ \text{grenzen}}} E_{p,q}\} \right.$$

Wir müssen aber auch in Betracht ziehen, dass es energetisch günstiger sein könnte i und j nicht paaren zu lassen, selbst wenn sie von den Paarungsregeln her dazu imstande wären. Die eigentliche Grösse, die wir minimieren wollen ist $F_{i,j}$, die freie Energie einer Sekundärstruktur auf dem Segment $[i, j]$, unabhängig davon ob i und j paaren:

$$F_{i,j} = \min\{E_{i,j}, \min_{i \leq h < j}(F_{i,h} + F_{h+1,j})\}$$

Die gesuchte Grösse ist $F_{1,n}$. Wenn wir sie ermittelt haben, verfahren wir wie oben und "backtracken" durch die $C_{i,j}$ und $F_{i,j}$ Tabellen und erhalten die Liste der Basenpaare, die zur freien Energie $F_{1,n}$ geführt haben.

In der Praxis muss man eine Reihe von Näherungen einführen, insbesondere weil es zu lange dauert nach allen möglichen multiloops zu suchen. Es gibt auch kaum zuverlässige Energiemessungen für die Vielfalt aller multiloops. Zur effizienten Suche von multiloops benötigt man mehr als nur die C und F Tabellen, aber das führt hier zu weit. Es ändert sich aber nichts am Prinzip.

Man darf nicht vergessen, dass die so ermittelten Sekundärstrukturen von einer mathematischen Minimierung resultieren. Das ist natürlich nicht der Faltungsprozess wie er *in vitro* geschieht. Es kann sein, dass Faltungstrukturen kinetisch bestimmte Strukturen sind und nicht notwendigerweise energetische Minima.

Das dynamic programming Prinzip kann auch ausgenützt werden, um die Zustandssumme Q über die möglichen Strukturen S, die eine Sequenz einnehmen kann, zu berechnen:

$$Q = \sum_{S} \exp(-F(S)/kT)$$

Hier resultiert die Zerlegbarkeit aus der mutiplikativen Zusammensetzung der Gesamtzustandssumme aus Teilzustandssummen auf Teilketten. Dies geht

natürlich wiederum auf die Additvität der Energien, die jetzt im Exponenten stehen, zurück. Die Zustandssume erlaubt die thermodynamischen Wahrscheinlichkeiten einer bestimmten Sekundärstruktur zu ermitteln, $P(S) = \exp(-F(S)/kT)/Q$. Das insofern wichtig als bei Raumtemperatur eine Sequenz nicht nur eine Struktur minimaler freier Energie einnehmen wird, sondern im Gleichgewicht mit einem Ensemble alternativer Strukturen ähnlicher Energie stehen wird.

Aus der Zustandssumme lassen sich eine Reihe von weitern thermodynamischen Grössen errechnen. Beispielsweise die spezifische Wärme C:

$$H = kT^2 \frac{\delta \ln Q}{\delta T}$$
$$C = \frac{dH}{dT}$$

Die Abhängigkeit $C(T)$ liefert die Schmelzkurve einer RNA.

7 Thermodynamik der Doppel-Helix Bildung

Um Sekundärstrukturen vorherzusagen, müssen wir die energetischen Beiträge der einzelnen Sekundärstruktur-Motive (Strukturelemente) kennen. Messungen werden üblicherweise an Oligonukleotiden durchgeführt. Meistens sind es optische Absorptionsmessungen oder kalorimetrische (differential scanning calorimetry) Bestimmungen.

Die Beiträge einer Helix oder einer Schleife hängen von der Sequenz ab. Es ist nicht möglich alle Sequenzvariationen eines Oligonukleotids, das beispielsweise 15 Nukleotide lang ist, zu messen. Um die freie Energie der Faltung *beliebiger* Sequenzen aus gezielten Messungen weniger Sonder-Sequenzen zu extrapolieren, hat sich in der Praxis das "Modell der nächsten Nachbarn" gut bewährt. Es besagt, dass sich die freie Bildungsenergie eines grösseren Strukturelements, etwa einer Helix, additiv aus den Beiträgen der Wechselwirkungen zwischen nächsten Nachbarn in einer Sequenz zusammensetzt. Im Fall der Helix wären das die aufeinanderfolgenden Paare (stacking) von Basenpaaren.

Das Modell der nächsten Nachbarn ist gerechtfertigt, weil ja der wesentliche Beitrag zur Stabilität der Struktur einer Polynukleotidsequenz vom Basen-

stacking benachbarter Basenpaare resultiert. Im Gegensatz dazu können hydrophobe Wechselwirkungen, die wesentlich zur Konformation von Proteinen beitragen, nicht als Wechselwirkungen zwischen nächsten Nachbarn aufgefasst werden. Es ist also die lokale Natur des Basen-stacking, das die Resultate der Messungen weniger Fälle auf beliebige Sequenzen anzuwenden gestattet und damit die Vorhersage von energetisch optimalen Sekundärstrukturen ermöglicht.

In einer antiparallelen Doppelhelix, die nur aus Watson-Crick Paaren besteht, gibt es 10 nächste Nachbarn. Warum? Es gibt 4 Basen, also 16 Paare aufeinanderfolgender Basen. Damit überzählen wir aber, denn: wenn wir $\mathrm{A}(5' \to 3')\mathrm{G}$ sagen, dann impliziert das schon ein anderes Paar, $\mathrm{C}(5' \to 3')\mathrm{T}$, am gegenüberliegenden Helixabschnitt, m.a.W.

$$\begin{array}{cc} \longleftarrow & \longleftarrow \\ \mathrm{GA} & \mathrm{TC} \\ \mathrm{CT} & \mathrm{AG} \\ \longrightarrow & \longrightarrow \end{array} =$$

Das ergibt $16/2 = 8$ Paare. Damit unterzählen wir aber, weil die soeben angestellte Betrachtung nicht für Palindorme gilt: $\mathrm{A}(5' \to 3')\mathrm{T}$ auf einem Strang impliziert dasselbe Paar $\mathrm{A}(5' \to 3')\mathrm{T}$ am anderen. Mit den Watson-Crick Regeln gibt es zwei nächste-Nachbarn-Palindrome, AT und GC. Also gibt es insgesamt $8 + 2 = 10$ mögliche nächste Nachbarn in einer antiparallelen Doppelhelix. Das heisst, dass eine Eigenschaft P einer Sequenz aus den Eigenschaften P_i $(1 \leq i \leq 10)$, die den nächsten-Nachbar Wechselwirkungen intrinsisch sind, und der relativen Häufigkeit χ_i dieser Wechselwirkungen zusammengesetzt werden kann[1].

Wie werden nun die thermodynamischen Beiträge solcher "Strukturatome" ermittelt? Hier soll das nur am Beispiel der Helix-Bildung erläutert werden. Im Wesentlichen versucht man die Gleichgewichtskonstante K für die He-

[1]Man braucht eigentlich weniger als 10 Parameter, denn es gelten Relationen zwischen den χ_i. So ist evidentermassen jeder Block aufeinanderfolgender A's durch ein nicht-A am $5'$ und $3'$ Ende begrenzt. Ergo:

$$\chi_{AT} + \chi_{AC} + \chi_{AG} = \chi_{TA} + \chi_{CA} + \chi_{GA}$$

Ebenso für G. Das bedeutet, dass wir in einem nächsten Nachbar Modell 8 unabhängige Parameter haben, d.h. 8 Grössen mit diesem Modell aus empirischen Daten herausholen können.

lixbildung zwischen zwei komplementären Oligonukleotiden zu bestimmen. Den Zusammenhang mit der Thermodynamik erhält man dann über die Beziehung $\Delta G = -RT \ln K$ (und weiter mit $\Delta H = -R[\delta(\ln K)/\delta(1/T)]_p$ und $\Delta S = (\Delta H - \Delta G)/T$).

Das Problem besteht darin die Gleichgewichtskonstante mit einer Messgrösse, wie etwa der optischen Absorption, in Zusammenhang zu bringen. Dazu benötigt man eine Modellvorstellung von der Helix-Bildung. Das einfachste (und auch in der Praxis am häufigsten verwendete) ist das sogenannte "Zwei-Zustände-Modell" oder auch "all-or-none"-Modell. Es wird angenommen, dass es nur zwei Kofigurationen für die komplementären Oligonukleotide gibt: entweder Knäuel-Zustand ("coil", "denatured") oder Helix ("native"). Nehmen wir einmal an unser Oligonukleotid, M, habe Länge N und sei ein Palindrom, also selbst-komplementär. Das vereinfacht die Diskussion, weil wir es nur mit einer molekularen Spezies zu tun haben. Im Zwei-Zustände-Modell bezeichne M_2 die Doppelhelix-Konfiguration:

$$M + M \leftrightarrow M_2$$

In der kinetischen Beschreibung dieser Reaktion wird zwischen einem Nukleationschritt, mit Gleichgewichtskonstante κ, und einem Elongationsschritt mit Gleichgewichtskonstante s unterschieden, wobei κ im Verhältnis zu s ausgedrückt wird: $\kappa = \sigma s$. Die Gleichgewichtskonstante der Reaktion ist dann einfach

$$K = \kappa \cdot s^{N-1} = \frac{[M_2]}{[M]^2} = \frac{X}{2(1-X)^2 \cdot C}, \tag{1}$$

wobei $C = 2[M_2] + [M]$ die Gesamtkonzentration an Oligonukleotid ist und X die relative Anzahl gepaarter Stränge bezeichnet: $X = 2[M_2]/C$.

Verfeinerungen des Zwei-Zustände-Modells berücksichtigen mehrere Konfigurationen zwischen getrennten Strängen und Doppelhelix. Jede dieser Konfigurationen i hat eine eigene Gleichgewichtskonstante K_i (für die Bildung aus getrennten Strängen). Die Gesamtreaktion von Einzelsträngen zu allen möglichen partiell gepaarten Konfigurationen hat die Zustandssumme q als Gleichgewichtskonstante, $q = \sum_i \exp(-\Delta G_i/RT) = \sum_i K_i$. Die verfeinerten Modelle unterscheiden sich in den zugelassenen Konfigurationen – "aligned model", "staggering zipper"usw.

Um an K heranzukommen muss man den Anteil an gebildeten Helices kennen. Den erhält man indem die Absorption (oder irgendeine Eigenschaft) A

misst, die sich am einfachsten linear zur relativen Zahl, X_b, an gepaarten Basen verhält. Dann haben wir:

$$X_b = \frac{A - A_s}{A_d - A_s},$$

wobei A_d und A_s die Absorptionen der Doppelhelix und des Einzelstranges sind.

Wenn das Zwei-Zustände-Modell gilt, dann ist in der Gleichung 1 $X = X_b$. (Wenn es nicht gilt, ist die Lage komplizierter, aber nicht hoffnungslos. Wir ersparen uns das.) Man misst nun X_b in Abhängigkeit der Temperatur T. Bei der Schmelztemperatur T_m ist im Zwei-Zustände-Modell $X = X_b = 1/2$. Eingesetzt liefert das

$$K(T_m) = \frac{1}{C} \qquad \text{(selbst-komplementäre Oligos)}$$

Man beachte, dass wenn wir nicht ein selbst-komplementäres Oligonukleotid nehmen, sondern zwei zueinander komplementäre Spezies I und J in *gleicher Konzentration* $[I] = [J]$, dann wird 1 zu:

$$K = \frac{[IJ]}{[I][J]} = \frac{2X}{(1-X)^2 \cdot C},$$

(X ist nun $[IJ]/([I] + [IJ])$ und $C = 2([IJ] + [I])$.) Bei Schmelztemperatur ergibt das

$$K(T_m) = \frac{4}{C} \qquad \text{(komplementäre Oligos).}$$

Eine ähnliche Bemerkung gilt für die Bildung einer Haarnadel-Helix, d.h. einer intramolekularen Helixbildung plus eine Schleife. (Wie lautet hier die Konzentrationsabhängigkeit der Gleichgewichtskonstanten?)

Wir erhalten zusammenfassend (für komplementäre Oligos):

$$-\ln K(T_m) = \ln \frac{C}{4} = \frac{\Delta G}{RT_m} = \frac{\Delta H}{RT} - \frac{\Delta S}{R}$$

d.h.

$$\frac{1}{T_m} = \frac{R}{\Delta H} \ln \frac{C}{4} + \frac{\Delta S}{\Delta H} \tag{2}$$

Das ist eine lineare Abhängigkeit der inversen Schmelztemperatur vom log der Oligonukleotidkonzentration und ist in der Praxis tatsächlich recht gut erfüllt. Daraus lassen sich die thermodynamischen Grössen für die Strukturbildung ermitteln. Um an die nächste-Nachbarn-Beiträge heranzukommen, braucht man nur noch Differenzmessungen anzustellen. Dabei vergleicht man die Messkurven 2 für zwei Situationen in denen sich die komplementären Oligos nur in der gewünschten Wechselwirkung unterscheiden. Der Abstand der beiden Kurven bei konstanter Schmelztemperatur ergibt das ΔG, das auf den Unterschied in den Oligos zurückzuführen ist. Bei gleichem Nukleationsschritt hängt dieses $\Delta\Delta G$ (die Differenz zweier freier Bildungsenergien) von den s zwischen nächsten Nachbarn ab (siehe 1). Z.B. Eine Messung der Differenz zwischen AAGCUU (plus Komplement) und AACGUU (plus Komplement) liefert

$$\Delta\Delta G = -RT \ln(s_{AG} s_{GC} s_{CU} / s_{AC} s_{CG} s_{GU})$$

Eine Reihe von Messungen mit geeigneten OLigos erlaubt die einzelnen s_i nächster Nachbarn aufzulösen und aus diesen die entsprechenden elementaren freien Bildungsenergien zu ermitteln. Diese Daten gehen in die energetische Sekundärstruktur-Optimierung ein.