Problemseminar "*Bio-Datenbanken*" im WS 2002/2003

Einführung
in
die Bioinformatik und
die Bio-Datenbanken

Betreuer: Dr. Dieter Sosna

Bearbeiter: Stephan Kühn

Inhalt

1. Einleitung

Die Molekularbiologie beschäftig sich mit Makromolekülen und deren Funktion bei der Regulation von biologischen Vorgängen. Von besonderem Interesse sind Moleküle wie DNA, RNA und Proteine.
Bei der Analyse dieser Makromoleküle und deren Funktion fallen viele Daten an.
Die Aufgabe der Bioinformatik ist nun Werkzeuge zur Verwaltung und Analyse dieser gewonnenen Daten zu liefern.
Gegenstand dieser Ausarbeitung ist die Einführung in die Grundlagen der Molekularbiologie und ein kleiner Überblick über den Einsatz von Datenbanken in der Molekularbiologie.

2. Wichtige Grundbausteine in der Molekularbiologie

Proteine (Eiweiße) sind Hauptbestandteile des Cytoplasmas. Sie sind Makromoleküle, die aus Aminosäuren durch Peptidbindungen entstehen.

2.1 Aminosäuren

Aminosäuren sind mehrfunktionelle organische Säuren, die wenigstens eine Carboxyl- und eine Aminogruppe enthalten. Alle Aminosäuren, die in Proteinen vorkommen, haben die gleiche Grundstruktur (Abb. 1). Sie bestehen aus einem zentralen Kohlenstoffatom, das man als α-C-Atom bezeichnet, an dem ein Wasserstoffatom, die Aminoguppe (NH_2),die Carboxylgruppe (COOH) und ein Rest gebunden sind. Sie unterscheiden sich nur im Aufbau des Restes.

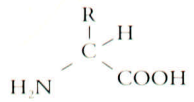

Abb. 1) Grundstruktur der Aminosäuren

Aminosäuren tragen sowohl die basische Aminogruppe wie auch die saure Carboxylgruppe, dadurch tritt ein Protonenübergang innerhalb des Moleküls auf (Abb.2).

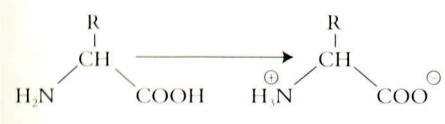

Abb. 2) Protonenübergang bei Aminosäuren

Deshalb tragen sie negative und positive Ladungen (Zwitterionen). Die Ladung des Moleküls kann sich durch Veränderung des pH-Wertes ändern.
Der isoelektrische Punkt ist der pH-Wert, an dem eine Aminosäure vollständig als Zwitterion vorliegt. Er hängt von der Säure- und Basenstärke der Aminosäure ab.
.

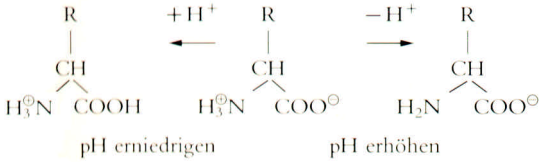

Abb. 3) Abh. der Ladung vom pH-Wert

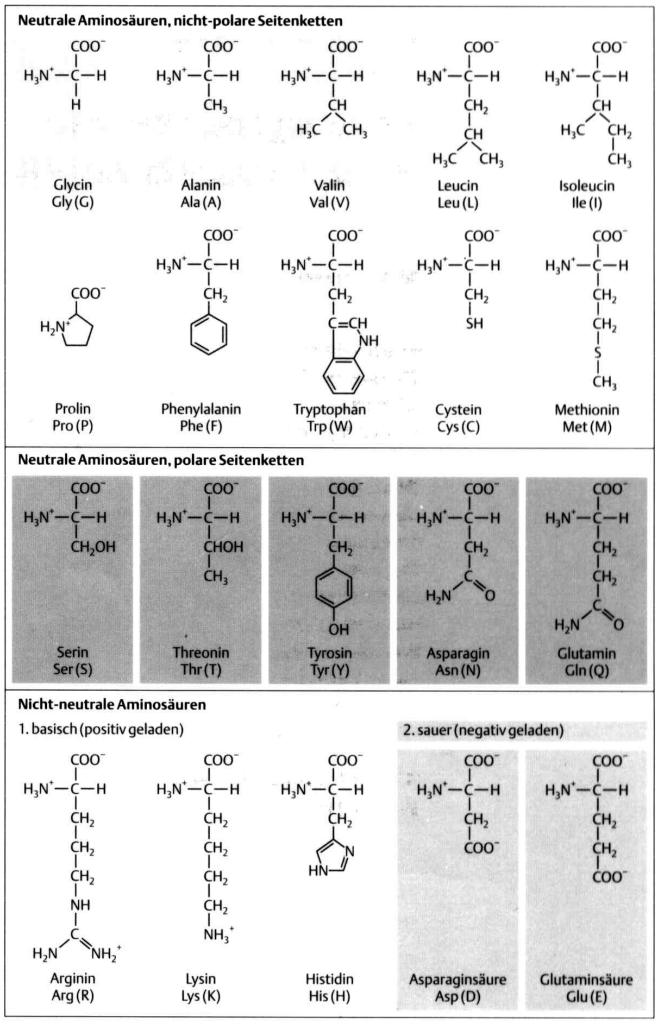

Abb. 4) Wichtige Aminosäuren mit ihrem Namen und Abkürzungen (im Drei- und Ein-Buchstaben-Code)

2.2 Peptidbindung

Die Carboxylgruppe einer Aminosäure kann sich mit der Aminogruppe einer anderen Aminosäure unter Austritt von Wasser verbinden (siehe Abb. 5) .
Die Peptidbindung ist infolge einer Elektronendelokalisation (Verschiebung der Bindungselektronen von Sauerstoff, Kohlenstoffund Stickstoff) eine Eineinhalbfach-Bindung. Dadurch bilden die einzelnen Atome einer Peptidbindung eine relativ starre Fläche (C- und N-Atom sind nicht gegeneinander drehbar) (siehe Abb. 6).

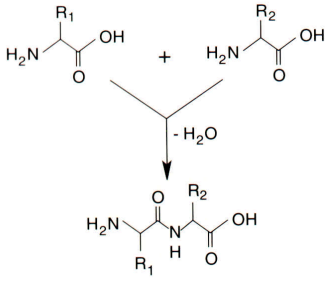

Abb. 5) Verbindung zweier Aminosäuren zu einem Dipeptid

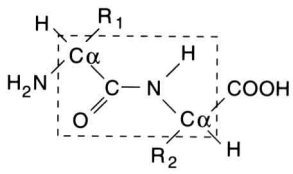

Abb. 6) Die Atome einer Piptidbindung liegen in einer Ebene

2.3 Peptide

Substanzen, die durch Verbindung mehrerer Aminosäuren entstehen, nennt man Peptide.
Besteht ein Peptid aus zwei Aminosäuren so ist es ein Dipeptid, enthält es drei Aminosäuren so ist es ein Tripeptid usw., D.h. die Peptide werden nach der Zahl enthaltenen Aminosäuren, nicht nach der Zahl der Peptidbindungen bezeichnet.
Peptide mit zehn oder weniger Aminosäuren werden als Oligopeptide und solche mit mehr als zehn Aminosäuren werden als Polypeptide bezeichnet.
Peptide weisen eine Polarität auf, an einem Ende des Peptids sitzt eine Aminosäure mit freier Aminogruppe (N-terminale Aminosäure, dieses Ende wird durch ein H [von H_2N- herrührend]

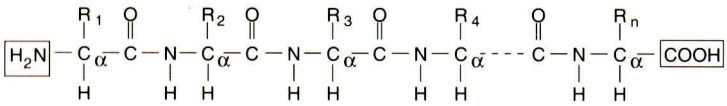

Abb. 7) Polypeptid

repräsentiert). Am anderen Ende sitzt eine Aminosäure mit einer freien Carboxylgruppe (C-terminale Aminosäure, symbolisiert durch ein -OH [von -COOH herrührend]).
Aminosäuren, deren Carboxylgruppe an dem Aufbau einer Peptidbindung beteiligt ist , erhält die Endung -yl, Z.B. Valyl-serin (H-Val-Ser-OH).

2.4 Proteine

Polypeptide mit mehr als 50-100 Aminosäuren nennt man Proteine.

2.4.1 Struktur der Proteine

Die Struktur der Proteine läßt sich in vier Strukturebenen gliedern:

- Primärstruktur: Aminosäuresequenz des jeweiligen Proteins
- Sekundärstruktur: Faltung der Polypeptidkette
- Tertiärstruktur: räumliche Struktur der Polypeptidkette
- Quartärstruktur: Anzahl und gegenseitige räumliche Anordnung der Untereinheiten

2.4.1.1 Die Primärstruktur der Proteine

In Proteinen treten 20 verschiedene Aminosäuren (proteinogene Aminosäuren) auf.
Die verschiedenen Proteine unterscheiden sich in Anzahl und Reihenfolge der Aminosäuren voneinander. Für Proteine, die aus 100 Aminosäuren aufgebaut sind, ergeben sich bereits 20^{100} Möglichkeiten die verschiedenen Aminosäuren anzuordnen.
Um ein Gefühl für die Größenordnung zu bekommen: 20^{100} bewegt sich in der Größenordnung von 10^{130}. Im Vergleich dazu hat unser Galaxie 10^{67} Atome.
So zum Beispiel unterscheidet sich das Insulin des Menschen nur in einer einzigen Aminosäure vom Insulin des Schweines.(am C-Terminus der B-Kette hat das Schweineinsulin Ala, das menschliche Insulin hingegen Thr).

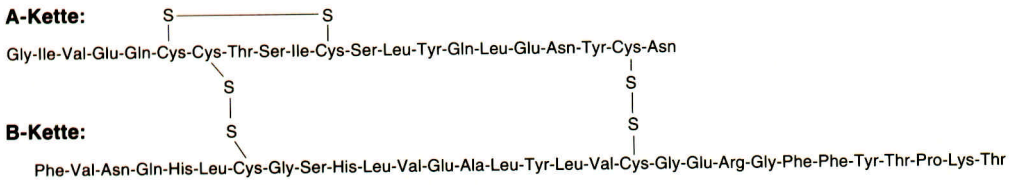

Abb. 8) Primärstruktur des Humaninsolins

Die Reihenfolge der Aminosäuren in einer Polypeptid Kette heißt **Primärstruktur**.

2.4.1.2. Die Sekundärstruktur der Proteine

Die Sekundärstruktur beschreibt die Faltung einer Polypeptidkette. Dies geht darauf zurück, dass die Peptidbindung eine ziemlich starre Ebene (siehe Kapitel 2.2) darstellt, so dass man sich die Polypeptidkette als eine Aufeinanderfolge starrer Ebenen vorstellen muss, die in einem bestimmten Winkel zueinander stehen.
Es gibt zwei Typen von Sekundärstrukturen: die Helix (Schraube) und das Faltblatt.

Die Helix

Die Helix kann rechts- oder linksgängig sein.
Rechtsgängigkeit herrscht, wenn die Schraube bei Blickrichtung vom N- zum C-Terminismus im Uhrzeigersinn verläuft und Linksgängigkeit, wenn sie diesem entgegenläuft.
Die Spirale in der Abb.9 ist rechtsgängig, da - wenn man hier von unten darauf schaut (Blickrichtung von N- nach C-Terminismus) - die Spirale im Uhrzeigersinn verläuft.

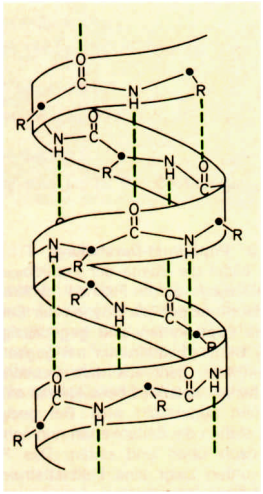

Abb. 9) Schematische Darstellung der α-Helix als Peptidkette

Die α-**Helix** ist eine stabile Sekundärstruktur.
Jede Windung hat 3,6 Aminosäuren und der Abstand zwischen zwei benachbarten Windungen beträgt 5,4 Å (Ångström,= 0,54nm), das ist die sogenannte Identitätsperiode.
Die NH- und CO-Gruppen stehen sich in diesem Abstand entlang der Hauptachse der Helix gegenüber. Wasserstoffbindungen bilden sich (von NH- zu OH-Gruppe) zwischen den einzelnen Windungen aus, d.h. sie verlaufen parallel zur Längsachse der Helix und verleihen ihr eine besondere Stabilität. Die Seitenketten der Aminosäuren stehen radial nach außen vom eigentlichen Schraubenkörper. Sie können miteinander oder mit dem Lösungsmittel in Reaktion treten.
Die Aminosäure Prolin ist mit der Regelmäßigkeit der Helix nicht vereinbar und führt zu einer Unterbrechung der Helix und einem "Knick" in der Polypeptidkette .
Wo Prolin in der Sequenz vorkommt, ergibt sich ein Abweichung von der regelmäßigen Struktur.
Das Haar- und Wollprotein Keratin ist ein Beispiel für ein Protein, das spontan eine α-Helix ausbildet.

Die Faltblattstruktur

Bei dieser Struktur kann man sich die Ebenen der Peptidbindungen als Seiten eines Endloscomputerpapiers, die in einem bestimmten Winkel aufeinander treffen, vorstellen. Wenn zwei solche Faltblätter nebeneinander liegen, entsteht die Möglichkeit der Ausbildung von Wasserstoffbrücken zwischen ihnen, die hier im Gegensatz zur α-Helix senkrecht zur Längsrichtung der Polypeptidketten verlaufen. Die Seitenketten der Aminosäuren stehen nahezu senkrecht zur Längsrichtung nach oben bzw. unten.

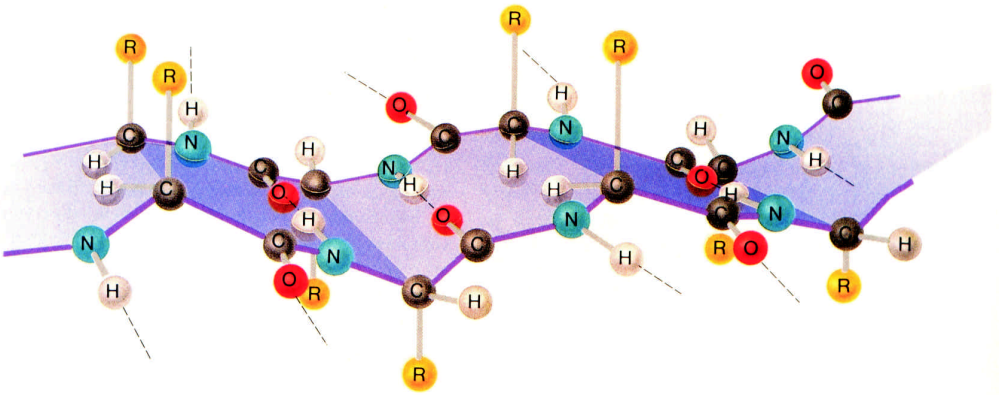

Abb. 10) Schematische Darstellung der Faltblattstruktur
Wasserstoffbrücken, hier gestrichelt dargestellt, verbinden die gefalteten Ketten

2.4.1.3 Die Tertiärstruktur der Proteine

Die globulären (kugelförmigen) Proteine besitzen im Gegensatz zu den fibrillären (fadenförmigen) Proteinen, die nur die Primär- und Sekundärstruktur aufweisen, eine räumliche Struktur - die Tertiärstruktur. Sie kommt durch die dreidimensionale Anordnung ihrer Polypeptidkette zustande.
Die Tertiärstruktur kommen drei Strukturelemente vor:

1. Helices (auch als α-Strukturen bezeichnet),
2. Faltblätter (auch als β-Strukturen bezeichnet) und
3. unregelmäßige Schleifen, die häufig auch als Haarnadelbiegungen ausgebildet sind.

Oft lässt sich die Ausbildung von Strukturdomänen beobachten. Strukturdomänen sind Teile der Tertiärstruktur, die bei miteinander verwanden Proteinen als Strukturelemente immer wiederkehren, eine bestimmte Anzahl von Helices, Faltblättern und Schleifen in bestimmter räumlicher Anordnung zueinander enthalten und durch bestimmte Funktionen gekennzeichnet sind.

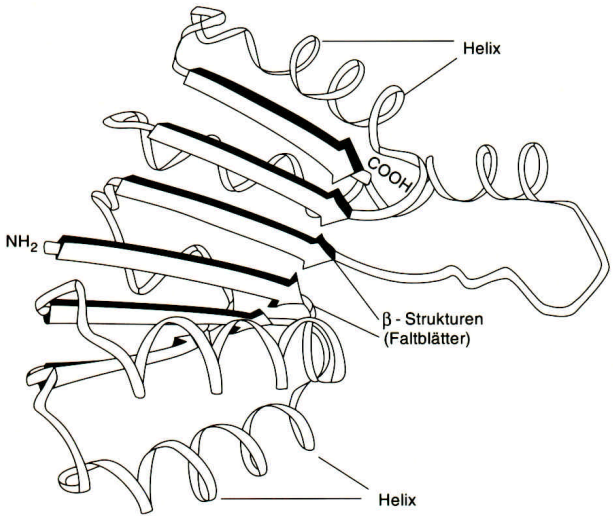

Abb. 11) Eine Strukturdomäne in der Raumstruktur eines Proteins

2.4.1.4 Die Quartärstruktur der Proteine

Wenn ein Protein aus mehreren Polypeptidketten, also aus Untereinheiten, aufgebaut ist, spricht man von einer Quartärstruktur. Man bezeichnet dieses dann auch als oligomeres Protein. Bevorzugt bestehen solche Proteine entweder aus vier Untereinheiten, dann nennt man ein solches Protein tetramer, oder aus zwei Untereinheiten und man spricht von einem dimeren Protein. Ein Protein, das nur eine Polypeptidkette enthält, bezeichnet man als monomer. Untereinheiten können bei oligomeren Proteinen entweder identisch oder nichtidentisch sein. Hämoglobin, zum Beispiel, ist ein tetrameres Protein. Es besteht aus 4 Untereinheiten die paarweise identisch sind und als α- und β-Ketten bezeichnet werden.

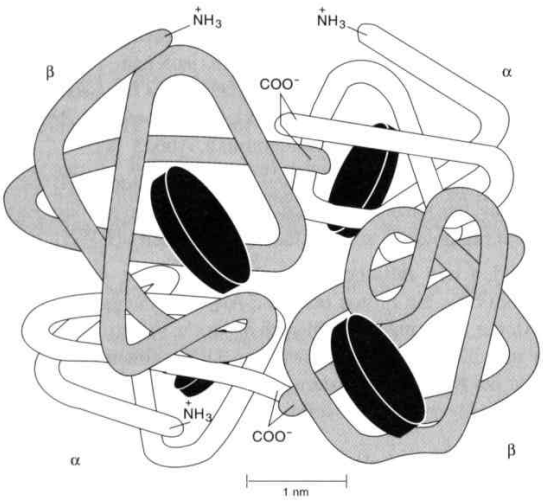

Abb. 12) Quartärstruktur des Hämoglobins $\alpha_2\beta_2$
Es sind jeweils die N- und C-Termini zu erkennen die α-Ketten sind weiß und die β-Ketten grau

2.5 Nucleotide und Nucleinsäuren

2.5.1 Die Bausteine

Nucleotide und Nucleinsäuren setzen sich aus drei verschiedenen Gruppen von Bausteinen zusammen:

- Purin- und Pyrimidinbasen
- Monosaccharide
- o-Phosphorsäure

2.5.1.1 Die Purin-, Pyrimidinbasen und Monosaccharide

In Nucleotiden und Nucleinsäuren findet man Adenin und Guanin als Purinbasen und Cytosin, Thymin, Uracil, 5-Methylcytosin (in der DNS von Bakteriophagen) als Pyrimidinbasen

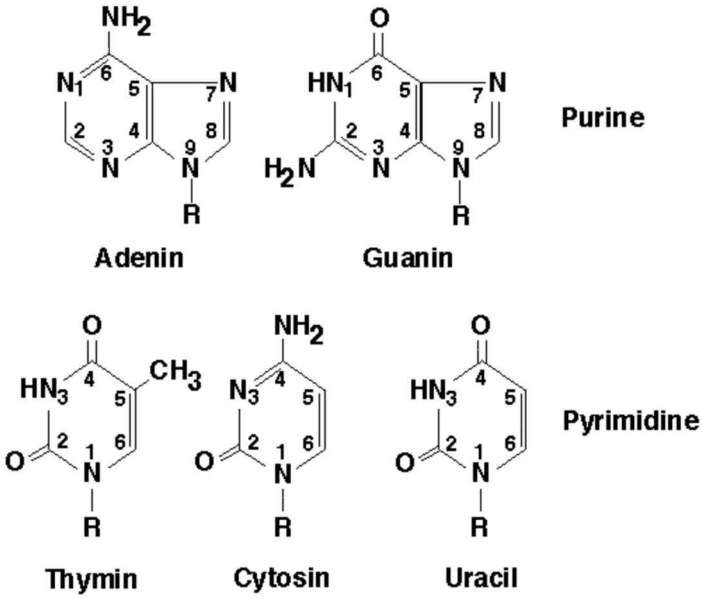

Abb. 13) Die Basen im Überblick, die Atome sind durchnumeriert

Die Monosaccharide sind D-Ribose und die 2-Desoxy-D-Ribose.
Sie enthalten fünf Kohlenstoffatome (man spricht auch von Pentosen).

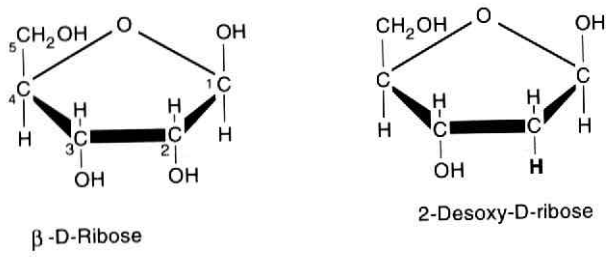

Abb. 14) die Pentosen, bei der ß-D-Ribose sind die C-Atome durchnumeriert, Numerierung erfolgt bei der 2-Desoxy-D-Ribose analog

2.5.1.2 Die Nucleoside

Ein Nucleosid ist die Verbindung zwischen einer Purin- bzw. Pyrimidinbase mit dem ersten Kohlenstoffatom (C_1) der D-Ribose.

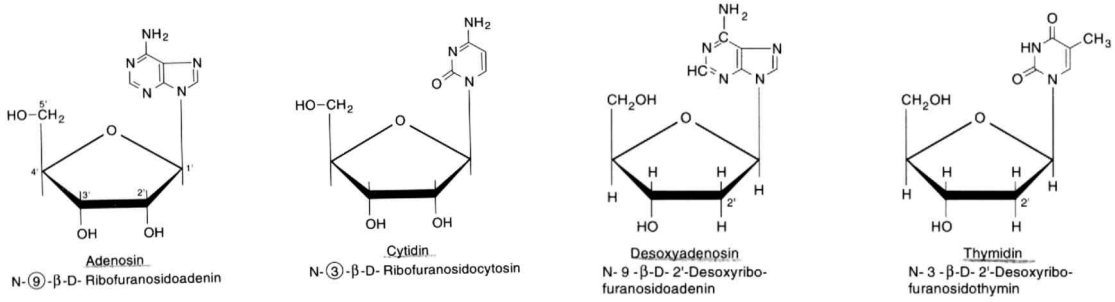

Abb. 15) Strukturformeln ausgesuchter Nucleoside

Verbindungen der Basen mit der 2-Desoxy-D-Ribose bezeichnet man als Desoxynucleoside.
Die Bindung erfolgt β-glycosidisch zwischen dem N-Atom 3 bei einer Pyrimidinbase bzw. dem N-Atom 9 bei einer Purinbase und dem C-Atom 1' der betreffenden Pentose.
Um Verwechslungen vorzubeugen, werden die C-Atome des Zuckers mit einem Strich versehen.
Die Nucleoside der Purinbasen werden gekennzeichnet durch die Endung -osin (Adenosin, Guanosin) und die Nucleoside der Pyrimidinbasen durch die Endung -idin (Cytidin, Uridin, Thymidin).

2.5.1.3. Die Nucleotide

Verbindungen der Nucleoside mit Phosphorsäure werden als Nucleotide bezeichnet.
Je nach beteiligtem Zucker unterscheidet man Ribo- und Desoxribonuleotide.
Die Phosphorsäure ist in diesen mit der Pentose esterförmig über die alkoholische OH-Gruppe am C-Atom 5' oder C-Atom 3' verbunden. Danach unterscheidet man 5'- von 3'-Nucleotiden.
Die Nomenklatur der Nucleotide berücksichtigt die Natur der Base und des Zuckers, sowie die Stellung des Phosphorsäuremoleküls am Zucker, zB: Adenosin-5'-monophosphat (auch als Adenylsäure bereichnet). Da Thymin in den natürlichen Nucleosiden und Nucleotiden nur in Verbindung mit Desoxyribose auftritt, verzichtet man bei diesen auf Kennzeichnung mit Desoxy.

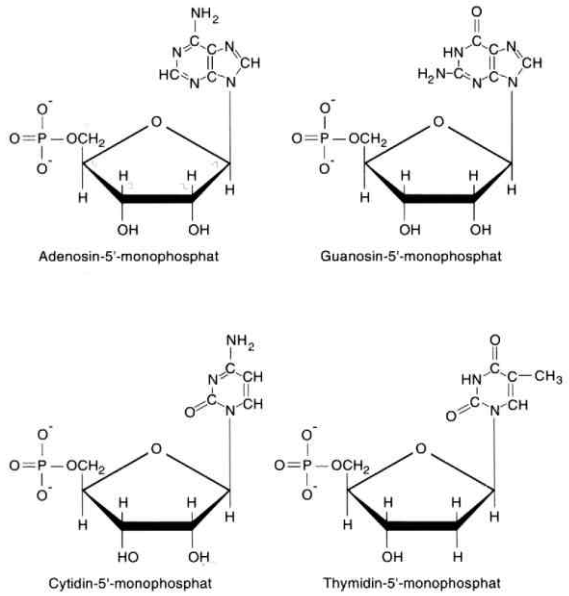

Abb. 16) Strukturformeln einiger Nucleotide

2.5.2 Die Nucleinsäuren

Nucleinsäuren sind Polynucleotide, d.h. sie setzen sich aus vielen Nucleotiden zusammen.
Ein einzelnes Nucleotid bezeichnet man als Mononucleotid.
Ihre Verknüpfung in den Nucleinsäuren erfolgt durch zweifach veresterte Phosphorsäure (Phosphorsäurediester-Bindung) zwischen dem C-Atom 5' eines Nucleotids mit dem C-Atom 3’ des benachbarten Nucleotids.
Zum Beispiel ergibt sich folgende Polynucleotidstruktur wie in der Abb. 17 und Abb. 18

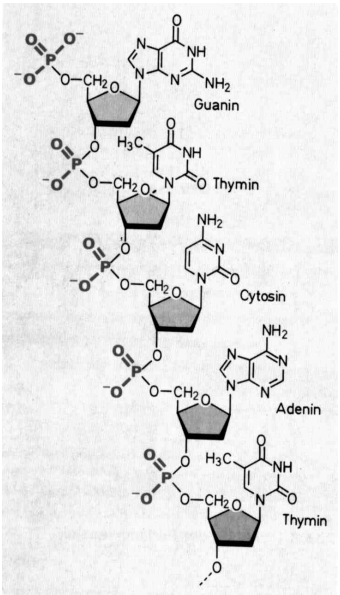

Abb. 17) Formelausschnitt einer DNA mit Phosphat am 5’-Ende

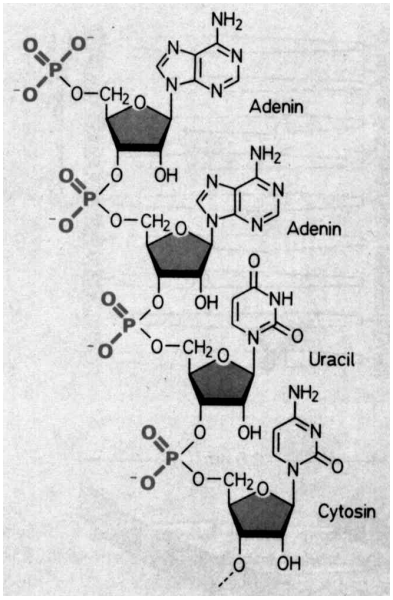

Abb. 18) Formelausschnitt einer RNA mit Phosphat am 5’-Ende

2.5.2.1 DNA und RNA

Je nach Art der Pentose unterscheidet man Nucleinsäuren. Ist der Zucker 2-Desoxy-D-Ribose spricht man von Desoxyribonucleinsäure (DNS) bzw desoxyribonucleic acid (DNA) und von Ribonucleinsäure (RNS) bzw. ribonucleic acid (RNA), wenn Zucker eine D-Ribose ist.
Ein weiterer Unterschied ist die Basenzusammensetzung:

In der DNA treten als Basen Adenin, Guanin, Cytosin und Thymin auf,
in der RNA kommt statt Thymin die Base Uracil vor.

2.5.2.2 Die Polarität der Nucleinsäurestränge

Nucleinsäuremoleküle besitzen eine Polarität. Sie kommt zustande, weil die Phosphorsäure nur die Hydroxyle 3' und 5' miteinander verknüpfen kann, da das C-Atom 1' der Pentose mit der Base und das C-Atom 4' mit dem furanoiden Ring besetzt ist. Dadurch ergeben sich zwei definierte Enden des Nucleinsäure-Moleküls. Nach Konvention schreibt man die Kette so, dass das 5'-OH-Ende, das noch einen Phosphat-Rest trägt, links und das 3'-OH-Ende rechts steht.
Die Kurzschreibweise der DNA-Struktur von Abb. 17 ist pdG-dT-dC-dA-dT oder d(pGpTpCpApT) und die Kurzschreibweise der RNA-Struktur aus Abb. 18 ist -A-A-U-C oder pApApUpC.

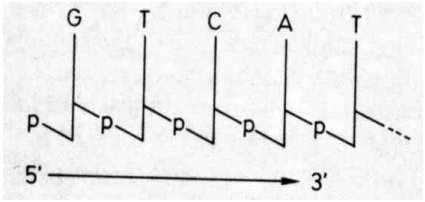

Abb. 20) Demonstration der Polarität, an dem DNA-Ausschnitt von Abb.17

2.5.2.3 Die Struktur der DNS

Die DNA tritt vorwiegend in Form von Doppelsträngen auf und besitzt eine
charakteristische Sekundärstruktur, die Doppelhelix.
Man kann sie wie folgt beschreiben:

1. Die zwei Polynucleotidstränge sind in der Doppelhelix spiralförmig verdrillt, sie müssen entdrillt werden, wenn sie getrennt werden sollen.
2. Die Polypeptidketten besitzen entgegengesetzte Polarität, sie verlaufen antiparallel
3. Es stehen sich immer zwei Basen, die strukturell komplementär sind, in der Doppelhelix gegenüber.

Man kann sich also die DNA als "verdrillte Strickleiter" vorstellen, wobei eine Windung 10 Basenpaare enthält und die Identitätsperiode 3,4 nm ist. Die komplementären Basenpaare sind Adenin - Thymin und Cytosin - Guanin. Zwischen den komplementären Basen bilden sich Wasserstoffbindungen aus (Zwischen Adenin und Thymin zwei und zwischen Guanin und Cytosin drei) . Diese Wasserstoffbindungen halten die Einzelstränge zusammen und stabilisieren so die DNA-Doppelhelix. Durch die Basenpaarung bestimmt jede Base ihren Partner, so dass ein Strang die vollständige Sequenz der Basen des anderen Strangs festlegt.
Die Zuckerreste stehen sich nicht diametral gegenüber. Dies bewirkt, dass die Windungen der Helices zum Teil weiter entfernt sind, zum Teil näher beieinanderliegen. Dadurch gibt es eine große und eine kleine Furche.

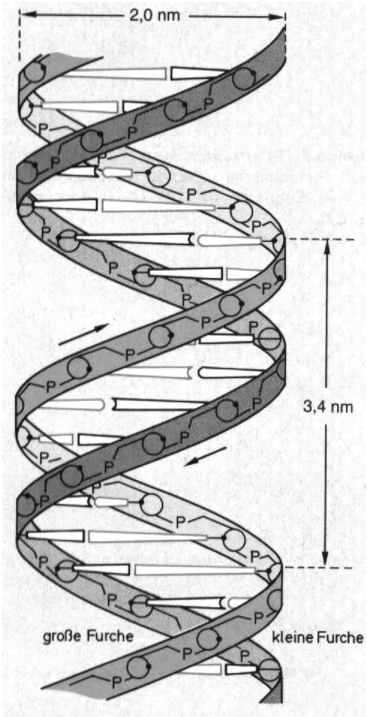

Abb. 21) zeigt die räumliche Struktur der DNA-Doppelhelix

Diese Grundform der DNA bezeichnet man als **B-DNA**. Die DNA kann noch in zwei weiteren Formen Vorliegen:
A-DNA ist bei geringer Hydratisierung zu beobachten.
Hier erscheinen die Basen geneigt gegenüber dem Zucker-Phosphat-Gerüst und eine vollständige Windung umfaßt 11 Basenpaare.

Bei der Z-Konformation der DNA ist die Schraubenrichtung entgegengesetzt zur B-Form. Sie hat diesen Namen erhalten, weil hier das Zucker-Phosphat-Gerüst Zickzack-förmig geknickt vorliegt. Einzelne Abschnitte können in Eukaryonten-Chromosomen in Z-Konformation vorliegen.
In Sequenzen in den sich die Purin- und Pyrimidin-Basen abwechseln kann es zu Bildung von Z-Abschnitten kommen, wenn die umgebende DNA durch ihre Überspiralisierung einen Zwang auf eine solche DNA ausübt.

3. Der genetische Code

Ein Gen ist molekulargenetisch betrachtet, ein Stück einer DNA-Doppelhelix, das in seiner Nucleotidsequenz die Information für ein Polypeptid enthält und daher eine funktionelle Einheit ist.
In den Proteinen der Lebewesen treten in der Regel 20 verschiedene Aminosäuren auf.
Deren Reihenfolge muss in der Nucleotidsequenz der DNA verschlüsselt vorliegen.
Wie wir in den vorherigen Kapiteln gelernt haben, kommen in den Nucleinsäuren 4 Basen vor. Würde eine Aminosäure durch eine Base bestimmt, so ließen sich den 4 Basen nur 4 Aminosäuren zuordnen. Zwei Nucleotide würden auch nicht ausreichen, denn dann könnten erst $4^2 = 16$ Aminosäuren durch die DNA codiert werden. Erst bei einer Kombination von 3 aufeinander-folgenden Nucleotide, genannt Nucleotidtriplett, ergeben sich genügend viele Möglichkeiten die 20 Aminosäuren zu codieren. Daraus folgt, dass es $4^3 = 64$ verschiedene Möglichkeiten von Triplettstrukturen gibt, d.h. wesentlich mehr als zur Verschlüsselung von 20 Aminosäuren nötig sind. Wichtig ist nun, dass alle denkbaren 64 Tripletts Codierungsfunktion besitzen und kein einziges Triplett überflüssig ist. Durch 61 Codons werden 20 Aminosäuren codiert, d.h. eine Aminosäure wird durch mehr als ein Codon Verschlüsselt. Deshalb nennt man den genetischen Code degeneriert. Die restlichen drei Codons besitzen Signalfunktionen. Ihre Anwesenheit markiert das Ende der Strukturinformation. Man nennt sie Stopp-Codons. Die Basen-Tripletts der DNA, die eine Aminosäure codieren, werden als Codogene bezeichnet. Dem Codogen entspricht nach der Transkription ein Codon auf der mRNA.
In der Code-Sonne (Abb.22) und Code-Tabelle (Abb. 23) sind die Codons der mRNA angegeben. Ihre Gesamtheit ist der genetische Code, und die Einheit der genetischen Information ist das Codon.
Wichtige Besonderheiten des Aminosäurecodes:

1. die Nucleotidtripletts werden nacheinander "abgelesen", jedes Triplett ist eine eigenständige Einheit, Teile von ihm gehören weder zum vorhergehenden noch zum nächsten Triplett
2. im genetischen Code gibt es keine Trennzeichen, d.h. die Tripletts folgen unmittelbar aufeinander
3. der genetische Code ist universell: von ganz wenigen Ausnahmen abgesehen, bedienen sich alle Organismen dieser Erde dieses Codes, d.h. sie nutzen die gleichen Codons für die gleichen Aminosäuren.

Von allen diesen allgemeinen Regeln gibt es einige interessante Ausnahmen:

1. Einige Bakterienvieren (Bakteriophagen) enthalten überlappende Gene mit verschiedenen Leserastern .Zum Beispiel enthält der Bakteriophage X174 zwei Gene mit eigenem Leseraster, die in größeren Genen mit anderen Leserastern enthalten sind.
2. Es gibt auch Ausnahmen bezüglich der Universalität des genetischen Codes: es wurde gezeigt, dass Mitochondrien für bestimmte Aminosäuren andere Codons verwenden. z.B.: UGA für Trp, AUA für Met sowei UAA und UAG für Gln.

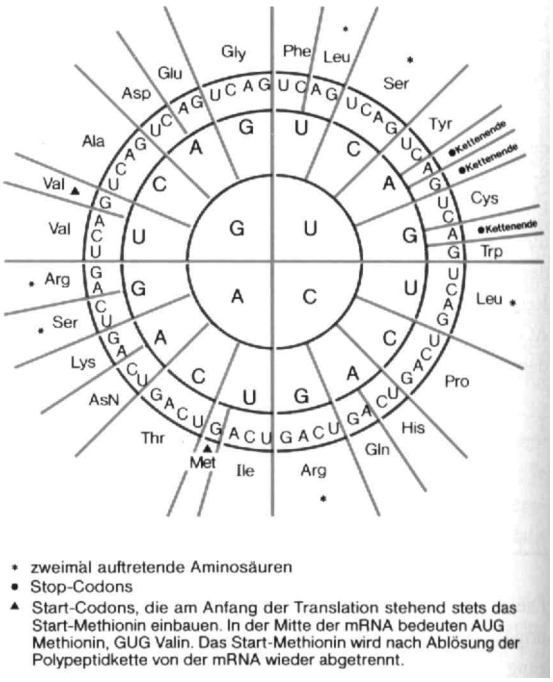

Abb. 22)Der genetische Code
Die Codewörter sind für die mRNA gegeben
Die Codons sind von innen nach außen zu lesen.

erste Position (5'-Ende)	zweite Position				dritte Position (3'-Ende)
	U(A)	C(G)	A(T)	G(C)	
U(A)	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr Stopp Stopp	Cys Cys Stopp Trp	U(A) C(G) A(T) G(C)
C(G)	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U(A) C(G) A(T) G(C)
A(T)	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U(A) C(G) A(T) G(C)
G(C)	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U(A) C(G) A(T) G(C)

Abb. 23)Der genetische Code in RNA- und in Klammern DNA-Schrift

4. Transkription

Definition:

Der Vorgang der Umschreibung der DNA- in die RNA-Schrift heißt Transkription.

Die Transkription lässt sich in drei Phasen Unterteilen: die Initiation, die Elongation und die Termination. Bei Eukaryonten schließt sich häufig noch eine Phase der Reifung (Processing, Prozessierung) der primären Transkripte an.

4.1 Transkription bei Prokaryonten

E.coli besitzt eine einzige DNA-abhängige RNA-Polymerase, die die Synthese aller RNA-Typen in diesen Bakterien katalysiert.

Initiation:

Betrachte nun den codogenen Strang.
Die RNA-Polymerase lagert sich an die Startstellen der Transkription, der sog. Promotor-Region, der DNA an. Diese Promotor-Region entspricht meist der Basenfolge 5'-TATAAT-3' oder ähnlichen Sequenzen (Pribnow-Box) und liegt etwa 10 Nucleotide oberhalb der transkriptierten Region bzw. des mRNA-Starts.

Elongation:

Nach der Initiation löst die RNA-Polymerase die Wasserstoff-Bindungen um den Startpunkt der mRNA und beginnt an der hier entwundenen DNA mit der Synthese.
Die RNA-Polymerase liest den Matrizenstrang ((-) Strang), dessen Polarität von 3' nach 5' gerichtet ist, ab. Der andere, nicht transkribierte Strang der DNA-Doppelhelix heißt codogener Strang ((+)Strang)).
Die RNA wird durch die RNA-Polymerase von 5' nach 3' synthetisiert. Die Synthese selbst beginnt mit ATP (Adenosintriphosphat) oder GTP (Guanosintriphosphat), an dessen 3'-OH-Gruppe die nachfolgenden Bausteine angelagert werden. Dadurch hat die neu gebildete RNA am 5'-Ende noch eine Triphosphatgruppe. Die Angliederung erfolgt komplementär zu den Basen des Matrizenstranges.

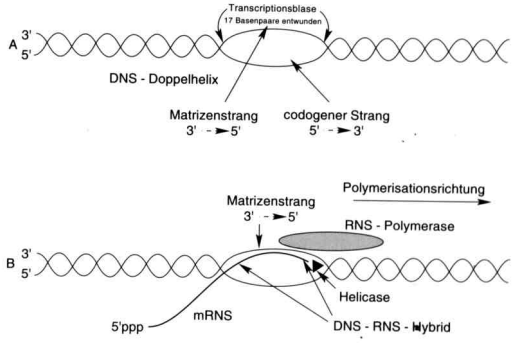

Abb. 24) Transkription
A: Codogener und Matrizenstrang, die Transkriptionsblase entsteht durch mehrere Enzyme und andere Proteine
B: RNA-Polymerase transkriptiert Matrizenstrang in der Taranskriptionsblase

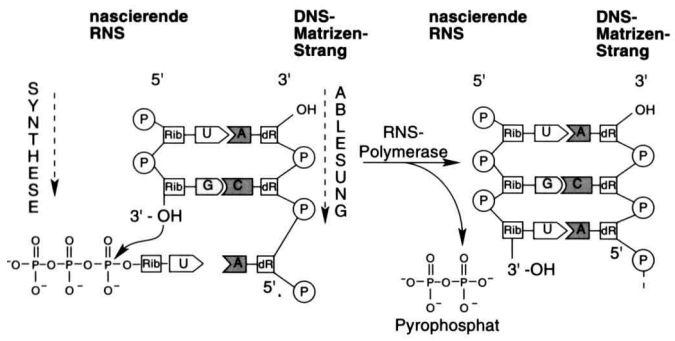

Abb. 25) Mechanismus der RNA-Polymerasereaktion (Rib. = Ribose; dR = Desoxyribose

Termination:

Bei Prokaryonten unterscheidet man zwei Klassen von Terminationssignalen, je nachdem ob diese einen Proteinfaktor, den sog. Faktor rho, brauchen.
rho-unabhängig ist eine RNA-Haarnadelschleife, die aus komplementären Basen besteht und sich spontan während der Transkription ausbildet. Ihr Zentrum liegt ca. 20 Nucleotide von RNA-Ende entfernt. Diesem folgen bis zum Ende des RNA-Transkripts Uridylsäurereste. Diese Haarnadelstruktur zerstört offenbar die RNA-DNA-Wechselwirkungen, so dass sich die RNA von der DNA löst.
Die zweite Klasse bilden die rho-abhängigen Gene. Bei ihnen wird auch am Ende eine kurze Haarnadel in der RNA gebildet. Ist hexameres rho-Protein vorhanden, so bricht die RNA-Polymerase ihre Wirkung ab und verläßt die DNA-Matrize.

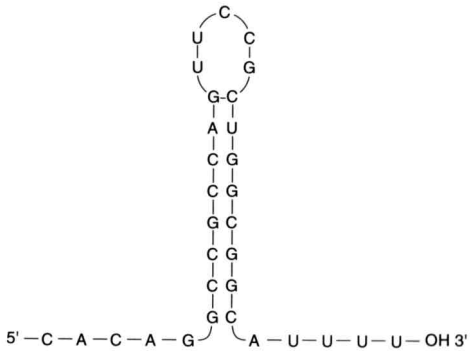

Abb. 26) Termination der RNA-Synthese
CG-reiche Haarnadelschleife mit vier terminalen Uracilresten am Ende der Synthetisierten Prä-RNA

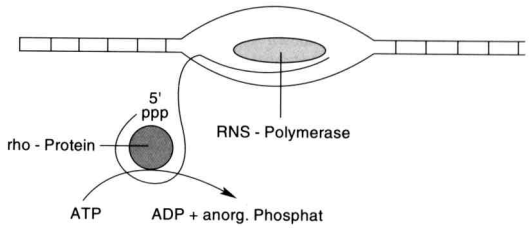

Abb. 27) Termination der RNA-Synthese
Rolle des rho-Proteins bei der Termination

4.2.Transkription bei Eukaryonten

Sie verläuft nach den gleichen Prinzipien, jedoch findet man im einzelnen einige Unterschiede. Während E.coli nur eine einzige RNA-Polymerase besitzt, kennt man bei Eukaryontenzellen drei verschiedene Typen des Enzyms, die als Polymerase I, II und III unterschieden werden. Die in den Nucleolus lokalisierte **RNA-Polymerase I** synthetisiert drei verschiedene ribosomalen RNAs (rRNAs). Im Nucleoplasma befindet sich RNA-Polymerase II und III.
Die **RNA-Polymerase II** transkriptiert die Gene für die Synthese der mRNA bzw. deren Vorläufer. Sie wird durch das Gift des Grünen Knollenblätterpilzes (α-Amanitin) stark gehemmt.
Bei hoher Konzentration hemmt dieses Gift auch die RNA-Polymerase III.
Die **RNA-Polymerase III** ist für die Synthese der kleineren RNA-Arten wie tRNA (transferRNA), ribosomalen 5S-RNA und einige kleine Kern-RNA-Arten (snRNA small nuclear RNA).
Auch bei den Eukaryonten steuern DNA-Abschnitte den Anfang der Transkription. Diese charakteristische Sequenzen sind Bindungsstellen von spezifischen Proteinen, die man als Transkriptionsfaktoren bezeichnet. Sie sind für die Initiation der Transkription verantwortlich, in dem sie die Anbindung der RNA-Polymerase an den Promotor ermöglichen.
Bei den Protein codierenden Strukturgenen, die von der RNA-Polymerase II transkriptiert werden, unterscheidet man drei verschiedene stromaufwärts vom Gen liegende Transkriptionsfaktoren bindende DNA-Segmenten mit jeweils spezifischen Consensussequenzen:
Die **TATA-Box**, das ist eine Adenin- und Thyminreiche Promotorsequenz mit typischer Folge "TATAAA", **CCAAT-** und **GC-Boxen**.
Ein weiterer Unterschied zwischen der Transkription bei Prokaryonten und Eukaryonten ist, dass die primären Produkte der Transkription bei Eukaryonten noch in vielfältiger Weise verändert werden. Diese Phase bezeichnet man als Reifung oder auch als Prozessieren (bzw. engl. processing) der RNA.

Processing:

Aus dem Transkriptionsprodukt, der Prä-mRNA (hnRNA), entsteht im Zellkern durch Umwandlung(Processing) die fertige mRNA. Dafür sind folgende chemische Modifikationen notwendig:

Die Umwandlung der Prä-mRNA beginnt damit, dass am 5'-Ende eine Cap-Struktur angehängt wird. Sie dient der Bindung der mRNA an das Ribosom zur Einleitung der Translation und als Schutz vor enzymatischem Abbau.

An das 3'-Ende der Prä-mRNA wird eine Poly-AMP-Sequenz angehängt, die etwa eine Länge von 100-200 Basen besitzt.

Die Vorstufen der mRNA-Moleküle enthalten Aminosäure codierende Sequenzen, die sog. Exons, und nichtcodierende, intervenierende Sequenzen, die Introns. Diese Instons werden herausgeschnitten und die übriggebliebenen Exons zusammen geführt. Diesen Vorgang nennt man Spleißen.

Nach dem Processing wird die reife mRNA vom Zellkern ins Zytoplasma transportiert, wo die Translation stattfindet.

rRNA- und tRNA-Synthese verlaufen nach dem selben Prinzip wie die mRNA-Synthese.

Der Ablauf der Transkription im Überblick:

1. Initiation:
   1.1 RNA- Polymerase (Enzymkomplex) bindet an einer best. Startsequenz der DNA an.
2. Elongation:
   2.1 Entspiralisierung und Aufspaltung der DNA-Doppelhelix an einer best. Stelle.
   2.2 RNA- Polymerase wandert in 3'-> 5'-Richtung auf dem Matrizenstrang ((-)Strang).
   2.3 Anlagerung von jeweils komplementären Ribonukleosidtriphosphaten nach dem Prinzip der Basenpaarung
   2.4 Verbindung der Ribonucleotide in 5'-> 3'-Richtung, dann Verdrängung des fertigen mRNA-Stranges und Wiederherstellung der DNA-Doppelhelix.
3. Termination:
   Beendigung der Transkription durch eine best. Terminationssequenz.
4. Processing/Reifung (bei Eukaryonten):
   Nichtfunktionsfähige Prä-mRNA umwandeln in die fertige mRNA

5. Translation

tRNA:

Die vier doppelsträngigen Bereiche und die drei einsträngigen Schleifen des tRNA-Moleküls bilden die typische Kleeblattstruktur.

Im Zentrum der 2. Schleife liegt das Anticodon, welches aus drei Basen besteht und durch die komplementäre Basenpaarung bestimmte mRNA-Codons erkennt. Die tRNA trägt am 3' (CCA'3) diejenige Aminosäure, die dem genetischen Code dem entsprechenden mRNA-Codon zugeordnet ist.

Viele Basen der tRNA sind chemisch verändert, zum Beispiel methyliert.

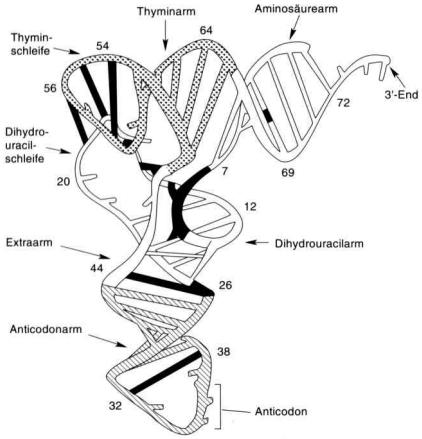

Abb. 28) Raumstruktur (Tertiärstruktur) eines tRNA-Moleküls

Ribosomen:

Ribosomen sind aus zwei Untereinheiten bestehende Zellorganellen, an denen die Proteinsynthese stattfindet. Sie sind im Zytoplasma, in den Mitochondrien und auf dem rauhen ER (Endoplasmatischem Reticulum) zu finden.
Ribosomen sind aus rRNA (ribosomaler RNA) und ribosomalen Proteinen aufgebaut und besitzen eine Durchmesser von 12 - 25 nm .
Die Ribosomen sind aus einer großen und einer kleinen Untereinheit aufgebaut.
Diese Untereinheiten werden jeweils nach ihrer Sedimentationskonstante K benannt.
Die Sedimentationskonstante wird in Svedberg-Einheiten (S) gemessen.
Ein vollständiges Ribosomen im Zytoplasma hat eine Sedimentationskoeffizienten von 80S, die Untereinheiten von 40S und 60S. Diese Werte werden in der Ultrazentrifuge ermittelt, sie lassen sich nicht addieren.
In Bakterien und Mitochondrien sind Ribosomen kleiner, ihre Sedimentationskonstante beträgt 70S, die der Untereinheiten 30S und 50S.
In Zellen mit hoher Wachstumsgeschwindigkeit (Tumorzellen, regenerierendes Gewebe) oder umfangreicher Proteinsysnthese (Leber, Nervenzellen) ist die Zahl der Ribosomen entsprechend hoch.
Im aktiven Zustand lagern sich Ribosomen entlang einem mRNA-Moleküls zu Polysomen zusammen.

Definition:

Die Übersetzung der in der Basensequenz der mRNA verschlüsselten Information in eine Aminosäuresequenz eines Proteins wird als **Translation** bezeichnet und läuft in den Ribosomen ab.

Die Translation lässt sich in fünf Phasen einteilen:
Aktivierung der Aminosäuren und deren Übertragung auf tRNA, Initiation, Elongation, Termination und Prozessierung

Aktivierung der Aminosäure und deren Übertragung auf tRNA

In dieser Phase wird die tRNA mit Aminosäure beladen. Dieser Vorgang verbraucht ATP zu AMP. Hier verbinden Enzyme die einzelnen Aminosäuren mit ihren entsprechenden tRNA-Molekülen. Diese Enzyme werden Aminoacyl-tRNA-Synthetasen genannt. Sie erkennen einerseits die zu übertragende Aminosäure und anderseits die zugehörige tRNA.
Eine tRNA, die eine Aminosäure trägt, nennt man Aminoacyl-tRNA.

Initiation:

Die Proteinsynthese beginnt mit der Bildung des ribosomalen Initiations- oder Startkomplexes. Dieser setzt sich aus folgenden Komponenten zusammen:

- der kleinen Untereinheit 40S eines Ribosoms
- der mRNA
- den Initiationsfaktoren
  bei Prokaryonten gibt es drei (IF-1 bis IF-3),
  bei Eukaryonten zehn (eIF-1 bis eIF-10)
- der Initiator-tRNS
- GTP
- Mg2+

Die Proteinsynthese beginnt stets mit einer Initiator-Aminosäure (Methionin, bei Bakterien mit N-Formylin). Die Ableserichtung der mRNA ist von 5' nach 3' und das Protein wird von seinem NH_2- zu seinem COOH-Ende hin synthetisiert. Das Startzeichen auf der mRNA ist ein Start- oder Inititator-Codon (5'-AUG-3' bzw. 5'-GUG-3', wobei AUG bevorzugt vorkommt). An dieses Startcodon wird eine Initiator-tRNA (tRNA, die die Initiator-Aminosäure trägt) über ihr Anticodon (3'-UAC-5') gebunden und gelangt so an die festgelegte Stelle im Startkomplex.
Eine purinreiche Sequenz, die stromaufwärts vom Startcodon liegt, ist komplementär zu einer bestimmten Sequenz am 3'Ende der 16S-rRNA der kleinen Untereinheit und trägt damit dazu bei, dass die mRNA in der richtigen Position im Initiationskomplex liegt.

Nachdem die Initiator-tRNA sich an die mit der kleinen Untereinheit verbundenen mRNA verbunden hat, erfolgt die Bindung der großen Untereinheit mit dem Startkomplex unter Freisetzung aller Initiationsfaktoren zu einem funktionsfähigen Apparat der Proteinsynthese. An diesem unterscheidet man zwei verschiedene Zentren:

- das Aminoacyl-tRNA-Bindungs- bzw. Acceptorzentrum (A-Zentrum)
- das Peptidyl-tRNS-Bindungszentrum (P-Zentrum)

Elongation:

In dieser Phase werden in Gegenwart von Elongationsfaktoren (EF) Aminosäuren Schrittweise verknüpft, so dass eine Polypeptidkette entsteht.

Die Initiator-tRNA (N-Formyl-Methionin-tRNAf) liegt am Anfang der Synthese im P-Zentrum, das A-Zentrum ist noch frei. In diesem befindet sich das zweite Codon, das direkt auf das Start-Codon folgt. An dieses bindet sich nun eine zu dem Codon passende Aminoacyl-tRNA (Aminosäure tragende tRNA), wodurch das A-Zentrum besetzt wird. Nun sind das P- und A-Zentrum des Ribosoms besetzt und die Bedingungen für die Bildung der ersten Peptidbindung zwischen der Initiator(fMethionyl-tRNAf) und der Aminoacyl-tRNA geschaffen.
Hier ist jetzt das Enzym Peptidyltransferase, die in der großen Untereinheit (S50/S60) sitzt. Die aktive Carboxylgruppe des Formylmethionins wird dabei mit der Aminogruppe der im A-Zenrum liegenden Aminoacyl-tRNA gebunden. Dabei entsteht eine Dipeptidyl-tRNA, die zunächst noch im A-Zentrum liegt.
Nun folgt ein Teilschritt, die Translocation. Hier bewegt sich das Ribosom, durch die Translocase katalysiert, um ein Triplett auf das 3'Ende der mRNA zu. Die dafür benötigte Energie wird aus der Spaltung von GTP zu GDP und Phosphat gewonnen. Die Dipeptidyl.tRNA wird von dem A-Zentrum in das P-Zentrum verschoben und gleichzeitig die entladene tRNA in das Zytosol/Zytoplasma verdrängt. Das A-Zentrum ist jetzt frei und kann die nächste beladene tRNA binden usw.

Die Verlängerung der Polypeptidkette nimmt demzufolge einen zyklischen Verlauf:

- Bindung einer Aminoacyl-tRNA an das A-Zentrum
- Knüpfung einer neuen Peptidbindung durch Übertragung des Peptidylrestes der Peptidyl-tRNA im P-Zentrum auf die Aminoacylgruppe im A-Zentrum
- Abgabe der freien tRNA aus dem P-Zentrum und vorrücken der Peptidyl-tRNA aus dem A-Zentrum in das P-Zentrum
- Bindung der nächsten Aminoacyl-tRNA an das freigewordene A-Zentrum

Termintion:

Die Polypeptidkettensynthese schreitet solange fort, bis eines der drei Terminations- oder Stoppcodons (UAG,UAA oder UGA) auf der mRNA das Ende signalisiert. Die Stopp-Tripletts codieren keine Aminosäure, d.h. es gibt in der Zelle keine tRNA, deren Anticodon komplementär zum Stoppcodon ist.
Zur Erkennung der Stoppcodons gibt es zwei Ablösungsfaktoren, die jeweils spezifische Stoppcodons im A-Zentrum erkennen und sich an sie binden und damit eine Wirkungsänderung der Peptidyltransferase bewirken. Die Peptidyltransferase verknüpft nun keine Peptide mehr, sondern spaltet die synthetisierte Polypeptidkette vom letzen Triplett ab und gibt sie frei. Ebenso verlassen auch die mRNA und tRNA das Ribosom, das nun wieder in eine große und kleine Untereinheit zerfällt und nun wieder zur erneuten Proteinsynthese zur Verfügung stehen.

Prozessierung:

Die Polypeptidkette wird nach der Termination durch irreversible chemische Vorgänge in ihre biologisch aktive Form umgewandelt. Man spricht hier von einer posttranslationalen Modifikation.

Die Wirksamkeit der Proteinsynthese wird gesteigert, indem sich mehrere Ribosomen gleichzeitig ein und dieselbe mRNA translatieren. Man nennt diese Aufreihung von Ribosomen, die von der mRNA zusammengehalten werden, Polysom.

6. Regulation der Genexpression

Das Jacob-Monod-Modell (Operon-Modell)

Zur Erklärung diese Modells der Regulation der Genaktivität beziehen wir uns auf Prokaryonten. Man kann die Gene je nach Funktion in drei verschiedene Gruppen einteilen:

Strukturgene, Operatorgene und Regulatorgene

Strukturgene sind verantwortlich für die Ausprägung von Eigenschaften. Ihre Nucleotidsequenz codiert die Aminosäuresequenz des Proteins/Enzyms.
Über diese Enzyme greifen sie in den Stoffwechsel ein.

Operatorgene kontrollieren die Funktionstüchtigkeit der Strukturproteine. Sie liegen in der DNA-Doppelhelix vor den Strukturgenen und können hemmend oder anregend auf die Transkription der Strukturgene wirken.

Als **Operon** bezeichnet man ein Operatorgen und die dazugehörigen Strukturgene, die bei einem Biosyntheseschritt zusammen wirken.

Der **Promotor** ist, wie wir im Kapitel der Transkription erfahren haben, eine Region des DNA-Stranges, an dem sich die RNA-Polymerase bindet.
Bei der Tanskription beginnt bei ihm die Synthese der mRNA. Der Promotor ist auch Bestandteil des Operons.

Regulatorgene steuern die Genaktivität, sie regulieren also die Operons. Dazu erzeugen sie spezielle Proteine, die sog. **Repressoren**. Diese Gene liegen außerhalb des Operons.

Die **Repressoren** lagern sich an die Operatorgene und verhindern so die Transkription der Strukturgene, d.h. das Operon ist inaktiviert.

Induktormoleküle verändern die Raumstruktur des Repressors, sie inaktivieren ihn, so dass er mit dem Operatorgen nicht mehr in Wechselwirkung treten kann und damit das Operon wieder aktiv wird und die Transkription ungehindert erfolgen kann.

Wenn das Substrat eines Enzyms des Operons die Genaktivität und damit die Enzymsynthese auslöst, spricht man auch von Substratinduktion.
Bei der Endprodukt-Repression oder Enzym-Repression hemmt ein Endprodukt einer Reaktionskette eine weitere Enzym-Synthese.

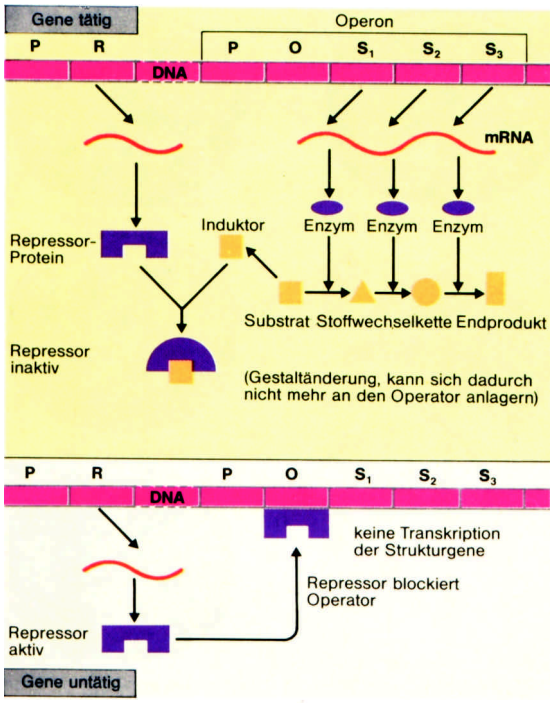

Abb. 29) Induktion
R Regulatorgen, O Operator, S Strukturgene, P Promotor

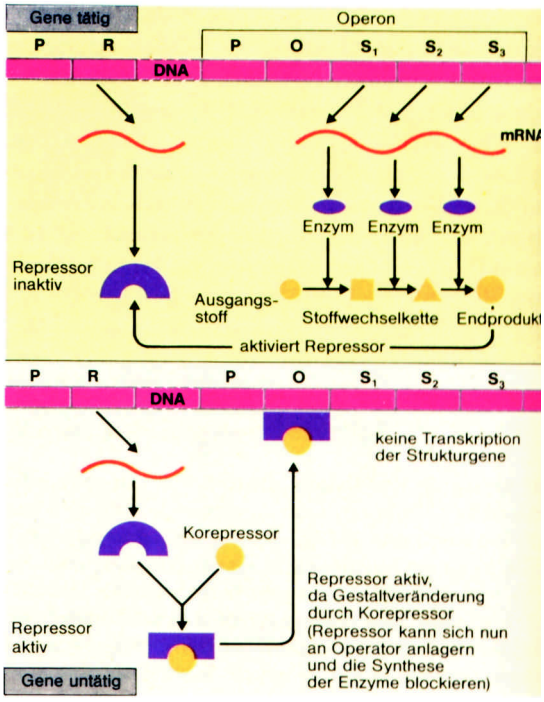

Abb. 30) Endprodukt-Repression
R Regulatorgen, O Operator, S Strukturgene, P Promotor

7. Datenbanken in der Bioinformatik

Die Bioinformatik beschäftigt sich mit Methoden zur Untersuchung von Problemen der Molekularbiologie auf Computersystemen. Auf Grund der dabei anfallenden sehr großen Datenmengen und umfangreichen Datenanalysen, sind Datenbanken in der Bioinformatik von großer Bedeutung.

7.1 Verwendung von Datenbanken in den verschiedenen Problemfeldern der Molekularbiologie:

Sequenz-Datenbanken zur DNA-Analyse und Sequenzierung

Ziel der Sequenzierung ist die Ermittlung der kodierenden und nicht-kodierenden Bereiche der DNA-Sequenz eines Organismus. Die durch die DNA-Analyse-Methoden erworbenen Daten werden in Sequenz-Datenbanken verwaltet. Diese Datenbanken sind meist sehr groß und besitzen exponentielles Wachstum.

Proteinsequenz- und Proteinstruktur-Datenbanken zur Strukturvorhersage

Ein Hauptziel der Biologie ist die Strukturvorhersage von Protein-Molekülen aus ihre Aminosäuresequenz, da die Funktion eines Proteins von seiner 3D-Struktur abhängig ist. Diese Vorhersage ermöglicht Laboruntersuchungen einzuschränken.
Bei Homologie-Basierten Ansätzen werden Aminosäuresequenzen bereits bekannter Proteine mit der Sequenz des unbekannten Proteins verglichen. Für diesen Zweck werden Proteinsequenz- und Proteinstruktur-Datenbanken aufgebaut und an diese Ähnlichkeitsanfragen gestellt.

Pathways/ Biochemische Pfade

Ein biochemischer Pfad modelliert abstrakt eine Abfolge von chemischen Reaktionen in einer Zelle. Metabolische Pfade (Reaktionswege im Stoffwechsel) und Regulatorische Pfade (Kontrollmechanismen in der Genexpression, siehe vorheriges Kapitel) sind von besonderem Interesse. Um biochemische Pfade zu finden, werden unter anderem Sequenz-Datenbanken genutzt. Die Verwaltung der dabei gewonnenen Daten geschieht ebenfalls in speziellen Datenbanken.

Sequenz-Datenbank zur Ermittlung phylogenetischer Bäume

Die Evolution geht mit Veränderung der Kodierung der Proteine der Organismen einher.
Um zu entscheiden, wann die Entwicklung der Nachfolger eines Organismus auseinander ging, werden spezielle Sequenzanalysealgorithmen, die auf Modellen der Geschwindigkeit der Veränderung der Kodierung der Proteine beruhen, angewandt.
Man erhält so die Stammbäume, die sog. phylogenetischen Bäume, der Organismen.
Bei der Berechnung dieser Bäume werden oft Sequenz-Datenbanken genutzt.

Genexpressionsanalyse

Ein Gen wird oft als DNA-Abschnitt definiert, der ein Protein kodiert. Die Transkription (siehe vorheriges Kapitel), insbesondere der Strukturgene, wird als Genexpression bezeichnet. Mittels sog. DNA-Chips kann das Expressionsniveau mehrerer tausend Gene gemessen werden, die die Zelle zu einem bestimmten Zeitpunkt exprimiert.
Gesunde und kranke Zellen können an Hand des Expressionsniveau unterschieden werden.
Data-Mining wird genutzt um Gene mit ähnlichen Expressionsmustern in Gruppen zusammen zu fassen.

7.2 Datenmodellierung und -management

Bei Bio-Datenbanken werden vier Formen von Datenmodellen verwendet:

- ASCII-Texte, sog. Flat-Files
- Datenmodelle von Standard Datenbanken,
  z.B.: relational, Objekt oder objektrelationale Datenbanken
- Object Protocol Model (OPM)
- ACEDB-Datenmodell.

Flat-Files

In der Anfangszeit der Molekularbiologie-Datenbanken wurden Datenbank Management Systeme (DBMS) wenig genutzt. Statt dessen wurden die Bio-Datenbanken aus indizierten ASCII Textdateien, den sog. "flat files", aufgebaut.
Einige Datenbanken basieren heute noch auf diesen. Ein Grund dafür ist, dass die Daten der Molekularbiologie oft sehr komplex sind, so dass viele flat-files tief geschachtelte Datensätze, Mengen, Listen und Abweichungen enthält, die sich nicht einfach in einem relationalen oder Objekt DBMS repräsentieren lassen.
Bio-Datenbanken, die als flat-files implementiert wurden, besitzen kein explizites Datenmodell. Ihre Einträge sind gewöhnlich implizit oder explizit durch Suchindexe strukturiert. Die meisten flat-file Sammlungen, die explizit strukturiert sind, nutzen Schlüsselworte, sog. "line types", als Index. Die Schlüsselworte und Indizes unterscheiden sich oft in verschiedenen flat-files nicht nur in ihrer Syntax, sondern auch von ihrer Semantik.
Viele Bio-Datenbanken, insbesondere Sequenzdatenbanken, sind noch als flat-file Sammlungen implementiert. Zudem sind heute flat-files der de facto Standard zum Datenaustausch in der Bio-Informatik. Und viele Werkzeuge in der Bio-Informatik (z.B.: BLAST und FASTA) arbeiten nur mit flat-files. Deshalb bieten viele Bio-Datenbanken ihren gesamten Inhalt in einem oder mehreren flat-files an.

Relationale und Objekt-Datenmodelle

Viele der auf flat-file beruhenden Datenbanken wurden auf relationalen, objekt-orientierten oder objekt-relationalen DBMS reimplementiert.
Ihre Daten werden folglich unter Verwendung des relationalen oder des Objekt- Datenmodels repräsentiert. Das Objekt-Modell ist besser geeignet als das relationale Model, um molekular-biologische Daten zu modellieren. Bio-Datenbanken basierend auf dem relationalen Modell besitzen oft sehr komplexe Schemas und sind damit nicht so intuitiv. Das macht es schwierig sie zu administrieren und Anfragen auf sie zu stellen.
Dennoch werden oft solche DBMS, wie z.B.: Oracle, Sybase der MySQL genutzt.

ACEDB

ACEDB ist ein DBMS, das ursprünglich für die Datenbank "A C.elegans Data Base" (abgekürzt ACeDB). Diese Datenbank enthielt Daten von einem kleinen Wurm "C.elegans". ACEDB ist eine Erweiterung dieses DBSM um andere solche spezialisierten Daten zu verwalten.
Bei ACEDB werden die Daten als Objekte, die in Klassen organisiert sind, modelliert. Jedoch unterstützt ACEDB weder Klassenhierarchie noch Vererbung.
Ein solches ACEDB-Objekt besitzt eine Menge von Attributen, die Objekte oder atomare Werte (Zahlen oder Zeichenketten) sein können.
ACEDB Objekte werden als Bäume, deren (beschrifteten) Knoten Objekte oder atomare Werte sind und deren Kanten Attributbeziehungen darstellen, repräsentiert.
Das Klassen-Modell von ACEDB spezifiziert die maximale Menge von Attributen, die ein Objekt der Klasse haben darf und den Typ/Classe dieser Attribute/Objekte.
ACEDB wird von Bio-Datenbanken oft zur Verwaltung genetischer Daten verwendet.
Der Quellcode von ACEDB ist public und kann deshalb den verschiedenen Anforderungen der speziellen Anwendungen angepasst werden.

Object-Protocol Model (OPM)

OPM wurde entwickelt zur Modellierung biologischer Daten und der Ereignisabfolge in wissenschaftlichen Experimenten. Es eignet sich gut zur Repräsentation zeitlicher Bedingungen und des Datenflusses zwischen den Teilexperimenten.
OPM eignet sich gut zur Modellierung dynamischer Daten, wie Phenotyp-Daten und die Dynamik von biologischen Prozessen.
OPM und dessen "data management tools suite" ist kommerziell.

Querverwiese

In Bio-Datenbanken verweisen Datensätze auf Beschreibungen der Experimente, durch die die Daten gewonnen wurden, auf ähnliche Daten in der selben oder einen anderen Datenbank. Meist werden diese Verweise durch (künstliche) Primärschlüssel realisiert und als Hyperlinks implementiert.
Diese Hypertext-Verlinkung ist ein besonders auffälliges Merkmal der Bio-Datenbanken.

Anfragen

Für Anfragestellungen bieten die Bio-Datenbanken oft Webformulare an.
Die derzeit genutzten Schnittstellen lassen sich leicht benutzen, es sind aber nur meist begrenzte Anfragen möglich und man kann dort Anfragesprachen im herkömmlichen Sinn nur selten finden. Einige Tools greifen direkt auf diese Schnittstellen zu. Zusätzlich stellen fast alle Bio-Datenbanken ihre Daten in den unterschiedlichsten Formaten als flat-files zum Download zur Verfügung.

7.3 Datenanalyse und -integration

Die meisten Bio-Datenbanken enthalten Software zur Datenanalyse. Diese Software sind entweder Implementierungen von den bekannten Bioinformatik-Algorithmen, wie zum Beispiel der Smith-Watermann-Algorithmus, oder Werkzeuge, wie BLAST, die auf bekannte oder wenige bekannte Algorithmen beruhen. Einige dieser Tools sind schwierig zu benutzen, da viele Parameter angegeben werden müssen und viele dieser Tools unzureichend dokumentiert sind.
Viele Bio-Datenbanken enthalten außerdem elementare Computer-Linguistik-Software zur Schlagwortsuche und Übersetzungen zwischen den geläufigsten Datenformaten.
Wegen des schnellen Wachstums von Bio-Datenbanken spielen Verfahren zur Knowledge Discovery und Data Mining immer größere Rolle.

Integration von Daten unterschiedlicher Ursprünge führt zu "Beschreibungs-", "Heterogenitäts-" und "semantischen" Konflikten.

Beschreibungskonflikt liegt vor, wenn das selbe semantische Objekt in verschiedenen Datenbanken unterschiedlich modelliert wird.

Heterogenitätskonflikt resultiert aus den unterschiedlichen Datenmodellen und Management-systemen der verschiedenen Datenbanken.

Semantischer Konflikt tritt auf, wenn die grundlegenden Begriffe, wie zum Beispiel "Gen" in verschiedenen Datenbanken unterschiedlich ausgelegt werden.

Frühe Werkzeuge zur Datenintegration berücksichtigen diese Konflikte nicht.
Neuere Ansätze versuchen Semantische Konflikte unter anderem mittels Ontologien zu lösen.
Das Problem, Daten unterschiedlicher Qualität zu integrieren, wurde bisher noch nicht zufriedenstellend gelöst.

Um die Daten aus den verschieden Datenbanken aktuell zu halten, sind regelmäßige (tägliche) Updates nötig. Bei Bio-Datenbanken ist dies besonders rechenintensiv, da flat-files der de facto Standard beim Datenaustausch sind. Strukturierte Modelle sind für den Datenaustausch vorzuziehen. Die semistrukturierte Herangehensweise zur Datenmodellierung und Datenmanagement scheint vielversprechend für die Datenintegration bei Bio-Datenbanken.
Verschiedene Forschungen beschäftigen sich mit der Modellierung molekularbiologischer Daten mittels XML.

8. Abschließende Bemerkung

Die Vorgänge, die in den Biowissenschaften (Molekularbiologie, Biochemie und Genetik) untersucht werden, sind sehr komplex und heute teilweise noch nicht bis ins letzte Detail bekannt. Bei der Erforschung dieser Vorgänge fallen große Datenmengen an, die mittels Methoden der Bioinformatik analysiert und verwaltet werden. Diese Methoden ermöglichen es in den Biowissenschaften Modelle leichter zu entwickeln und in diesen zu rechnen.
Das Potential der Informatik ist heute bei den Biowissenschaften bei weitem noch nicht ausgeschöpft. Somit ist die Bioinformatik ein ähnlich interessantes Forschungsgebiet, wie die einzelne Biowissenschaften selbst, in denen die Bioinformatik als Werkzeug eingesetzt wird.

Literaturverzeichnis

HO96 Prof. Dr. Hofmann, E.: *Medizinische Biochemie systematisch*
UNI-MED Verlag AG, 1996, ISBN 3-89599-121-X

KA88 Karlson, P.: *Kurzes Lehrbuch der Biochemie für Mediziner und Naturwissenschaftler*
Georg Thieme Verlag Stuttgart, New York, 1988, ISBN 3-13-357813-8

AB97 hrsg. von Abdolvahab-Emminger, H.: *Physikum EXAKT - Das gesamte Prüfungswissen in einem Band* Georg Thieme Verlag Stuttgart, New York, 1997, ISBN 3-13-107031-5

BK89 Bayrhuber,H. , Kull,U.: *Linder Biologie - Lehrbuch für die Oberstufe*
Schroedel Schulbuchverlag, 1989, ISBN 3-507-02347-4

MS88 Miram,W., Scharf,K.-H.: *Biologie heute SII - Neubearbeitung*
Schroedel Schulbuchverlag; 1988, ISBN 3-507-10540-3

FI92 Fink, E.: *Biologie* BON-MED, 1992, ISBN 3-928730-63-0

TR97 Trachsel, H. : *MOLEKULARBIOLOGIE - Repetitorium für Studierende der Human- und Veterinär-Medizin* http://ntbiouser.unibe.ch/trachsel/teaching/MBVorl/MBVorl.htm; 1997

FK02 Francois, B. , Kröger,P.: *A Computational Biology Database Digest:Data, Data Analysis, and Data Management* http://www.pms.informatik.uni-muenchen.de/publikationen/
2002

KF02 Francois, B. , Kröger,P.: *Aktuelles Schlagwort "Datenbanken in der Bioinformatik"*
http://www.pms.informatik.uni-muenchen.de/publikationen/, 2002

GI GI e.V.: *Informatik-Lexikon: Bioinformatik*
http://www.gi-ev.de/informatik/lexikon/inf-lex-bioinformatik.shtml

WM1 www.webmic.de:*Transkription*
http://www.webmic.de

WM2 www.webmic.de:*Proteinsynthese*
http://www.webmic.de

Bildverzeichnis

[BK89] Abb.1, Abb.2, Abb. 3, Abb. 10, Abb. 22, Abb.29, Abb. 30

[AB97] Abb.4

[HO96] Abb.5, Abb.6, Abb.7, Abb.8, Abb.11, Abb.12, Abb.14, Abb.15, Abb.16, Abb.23, Abb.24, Abb.25, Abb.26, Abb.27, Abb.28

[KA88] Abb.9, Abb.17, Abb.18, Abb.20, Abb.21