Bioinformatik für die Analyse mikrobieller Genome, Metagenome und Metatranskriptome

InterPro:

→ protein sequence analysis & classification

InterPro provides **functional analysis of proteins** by classifying them into **families** and predicting domains and important sites. We **combine protein signatures** from a number of member databases into **a single searchable resource**, capitalising on their individual strengths to produce a **powerful integrated database and diagnostic tool**.

InterPro combines signatures from **multiple, diverse databases** into **a single searchable resource**, reducing redundancy and helping users interpret their sequence analysis results. By uniting the member databases, InterPro capitalises on their individual strengths, producing a powerful diagnostic tool and integrated resource.

InterPro integrates **protein-family data** from **11 major sources**, classifying the different protein family, domain and functional site definitions hierarchically to **provide a unified view of diverse data**. **Pfam**, a member database of InterPro, generates new protein family entries and has the largest sequence coverage of any of the InterPro member databases. Both InterPro and Pfam have a number of important applications, including the automatic annotation of proteins for **UniProtKB/TrEMBL** and genome annotation projects. **Rfam** classifies non-coding RNA sequences into families, using probabilistic models that take into account both sequence and secondary structure information, termed covariance models (CMs). **Rfam** is uniquely placed to **annotate non-coding RNAs** in genome projects and is a major contributing database to **RNAcentral**, a sequence resource launched in **2014**.

When to use **InterPro**:

You can use **InterPro** if you have **an amino acid sequence** or a **set of sequences** and you want to know:

(a) what they are, what (b) **family** they belong to
(b) what is their **function** and (d) how it can be explained in structural terms

You can also use **InterPro** for a variety of other purposes, such as **examining the structural or functional predictions for any sequence already in the UniProt database**.

So erhalten wir bei einem **InterProt Ergebnis** zum Beispiel folgende Darstellungen:

GO term prediction
Biological Process

GO:0006955 immune response
Molecular Function

GO:0005125 cytokine activity
Cellular Component

GO:0005576 extracellular region

tRNA prediction:

→ **Transfer RNAs** (**tRNAs**) play an essential role in cell viability. Beyond their **major function** in translating the genetic code, tRNAs are implicated in many other processes such as **viral replication**, amino acid biosynthesis or cell wall remodeling.

→ The **DNA sequence** of **tRNA-genes** is **poorly conserved**, but **their secondary-structure is conserved**. This is the major reason why **tRNAscan-SE** works here.

Software includes:

→ **tRNAscan-SE** (http://lowelab.ucsc.edu/tRNAscan-SE/)

tRNAscan has **different modes**, such as **the search mode**: this one determines which probabilistic model to use in searches. Each model is based on **tRNA training data** from selected species or phylogenetic group. If no explicit model for the species of interest is available in the user interface, specifying either a general model or a model from a related species generally yields **good results**. Different search modes can offer varying speed and sensitivity. For **tRNAscan-SE**, the standard search modes are **fast** and very sensitive in most instances. However, for short sequences that contain extremely **atypical tRNAs**, the "**Infernal without HMM filter**" mode offers **slightly better sensitivity**.

Plasmide:

→ die **Umwelt kann variieren**. Plasmide sind prinzipiell **mobile Einheiten**, die einen Informationsaustausch zwischen individuellen Bakterien **ermöglichen**. Plasmide liegen oft in mehreren Kopien vor (z.B 5-15 pro bakterielle Zelle). Gene, die in großer Quantität benötigt werden, können so zum Beispiel parallel transkribiert werden.

Es gilt daher, das **Plasmide** eine Vorteil für einen Mikroorganismus darstellen, wie zum Beispiel durch:

(1) **Mobilität** (zum Beispiel Antibiotika-Resistenzen)

(2) **hohe Kopienzahl** (vorteilhaft)

Analytische Genomik: → dies umfasst die **Transkriptomik** und die **Proteomik**, inklusive **Massenspektrometrie**.

Barcoding: → Versehen mit unterschiedlichen **Adaptern**. ("Adapters may include sample specific tags." - Shendure J & Ji H. Nature Biotech 26, 2008)

Bakterielle Genome: → Das **Bakterium** *Dinoroseobacter shibae* (Rhodobacteraceae) hat ein **Genom** von (gerundet) 3.800.000 bp. Zudem hat es 5 Plasmide, die in ihrer Grösse zwischen ~72.000bp bis hin zu 190.000bp liegen. Archaea haben in ihrem Genom eher einen "Peak", Bakterien hingegen zwei peaks. Der Grund für Archaea ist wohl das sie für eine Nische adaptiert und optimiert sind; Bakterien hingegen, mit Ausnahme parasitischer Bakterien, in mehreren Nischen funktionieren müssen (Boden, Wasser → Schlamm → Fluss/See, etc...). Hier kommt auch die Konkurrenz zu anderen Arten zu tragen; siehe auch **Streptomyces** + **Isolierung von Antibiotika**.

Metagenomische Studien – Probleme und Einschränkungen zwecks Taxonomiebestimmungen:

Metagenomic sequencing provides a unique opportunity to explore earth's limitless environments harboring scores of yet unknown and mostly unculturable microbes and other organisms. Functional analysis of the metagenomic data plays a central role in projects aiming to explore the most essential questions in microbiology, namely:

'In a given environment, among the microbes present, what are they doing, and how are they doing it?'

Toward this goal, several large-scale metagenomic projects have been conducted or are currently underway.

Functional analysis of metagenomic data mainly suffers from the **vast amount of data generated** in these projects. The shear amount of data **requires much computational time and storage space**. These problems are compounded by other factors potentially affecting the functional analysis, including, **sample preparation**, **sequencing method** and **average genome size of the metagenomic samples**.

In addition, the read-lengths generated during sequencing influence sequence assembly, gene prediction and subsequently the functional analysis. The level of confidence for functional predictions increases with increasing read-length.

Usually, the most reliable functional annotations for metagenomic sequences are achieved using **homology-based approaches** against publicly available reference sequence databases. Here, we present an overview of the current state of functional analysis of metagenomic sequence data, bottlenecks frequently encountered and possible solutions in light of currently available resources and tools. Finally, we provide some examples of applications from recent metagenomic studies which have been successfully conducted in spite of the known difficulties. → https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3504928/

Each stage of the analysis suffers heavily due to inherent problems of the **metagenomic data** generated, including:

- **incomplete coverage**

- **massive volumes of raw sequence data produced by the next-generation sequencers**
- **generally short read-lengths**
- **species abundance and diversity**

Due to **shorter read-length** the overall **functional composition** is comparatively poor for shorter Illumina-sequencing derived reads than for longer Sanger reads.

Metagenomes/**Metatranscriptomes**: DNA and RNA from communities of uncultivated species, e.g.

→ **sea water**
→ **soil**
→ **sludge**
→ **human microbiomes**

→ In **Environmental Genomics**, we assemble "**partial genomes**".

Erstes Problem: wie gewinnt man die Genome dieses Organisme? Viele sind schwer erreichbar. Die wenigsten sind kultivierbar. Wenn wir mehr als 90% nicht kultivieren können, wie können wir die realen taxonomischen Verhältnisse wiedergeben? Selbst mit **PCR-Sampling** kann man ein *bias* nicht völlig ausschliessen.

Ein weiteres Problem ist, dass vielfach sogar geübte Taxonomen nicht in der Lage sind, Spezies mit der gebotenen Zuverlässigkeit zu bestimmen. Während größere Tiere und Pflanzen in der Regel sehr sicher bestimmt werden, ist die Zuordnung mikroskopisch kleiner Organismen sogar für Fachleute in vielen Fällen nicht mit 100%iger Genauigkeit möglich. So konnten in Tests geübte Personen Stichlinge mit einer Genauigkeit von 84 bis 95 Prozent bestimmen, bei Phytoplankton-Spezies sank die Treffsicherheit aber auf nur 72 Prozent.

Abhilfe könnten **bildbasierte automatisierte Identifizierungssysteme** schaffen, z. B. das Digital Automated Identification System (**DAISY**). So konnte **DAISY** 15 Spezies einer parasitischen Wespe mit 100%iger Genauigkeit anhand digitalisierter Bilder der Flügel bestimmen, wobei jede Identifizierung weniger als eine Sekunde benötigte.

Bei **Prokaryoten** kommt das Problem hinzu, das Gene transferiert werden können.

Prokaryotic taxonomy consists of **three separate components**: **classification** (i.e., the arrangement of organisms into groups or taxa), **nomenclature**, and **identification**. Although there is no official classification for prokaryotes, the classification system represented by **Bergey's Manual of Systematic Bacteriology** is widely accepted by the community of microbiologists and therefore is currently considered the best approximation to an official classification.

The **Bergey's** classification system is based on the phylogenetic analysis of the **small-subunit rRNA genes** (16S rRNA), as well as on **classical microscopic and biochemical observations** about the relatedness of the organisms, such as **G+C content deviation** and **DNA-DNA hybridization efficiency**. This system has been valuable in describing and appreciating the breadth of prokaryotic diversity and setting the framework for the study of relationships between taxa. Further, results from new approaches enabled by the availability of whole-genome sequences, such as phylogeny based on shared content of orthologous genes, indels, or signature sequences and concatenated alignments of many proteins, are generally congruent with the **16S rRNA gene-based phylogeny**, which adds further value to the system.

It is important to realize, however, that the definition or standards for the existing taxonomic ranks are far from being well delineated, particularly for the ranks higher than the species. In fact, considerable subjectivity in designating genera, families, etc., has been allowed, which is partially attributable to the great biochemical and morphological diversity exhibited by prokaryotes that prevents the employment of the same measuring rules for all groups of organisms.

Currently, the only major prerequisite for designating novel taxonomic ranks higher than the species rank is that clustering by 16S rRNA gene data should support such designations, but **no standards exist** in regard to the absolute differences between the taxonomic ranks. Consequently, the prokaryotic taxonomy represents, unavoidably, an artificial system, which often depends more on the intuition of individual researchers than on specific standards or knowledge of the natural history of organisms. Nonetheless, there is **great comparative value** in having a taxonomic system predictive of phenotypic and genetic relatedness of the grouped organisms and taxonomic ranks that are comparable, in terms of absolute differences and similarities, among lineages. It remains unclear, however, how the prokaryotic taxonomy is performing with regard to these issues, **partly due to the focus on the 16S rRNA gene**, which has overlooked the overall biochemical or genetic relatedness at the whole-cell level, and partly because of technological constraints in studying the differences and

similarities among microorganisms.

The recent availability of complete sequences of a number of prokaryotic genomes has made it possible to **study the genetic and functional relatedness between organisms at the whole-cell level**, and hence, to **provide novel insights** into the issues described above and an independent assessment of what the 16S rRNA-based system really represents.

However, **genomic studies** to date have mostly been focused on assessing the **accuracy** of phylogenetic reconstruction, particularly in the light of **horizontal gene transfer** (**HGT**), rather than the absolute differences between taxa and/or have failed to address the latter issue systematically for all prokaryotic taxa.

Faustregel: Wenn wir eine geringere Ähnlichkeit der 16sRNA Nukleotidsequenz als 97% haben, handelt es sich um eine neue Spezies.

Zwei Genome sollen auf "**Homologie**", "**Paralogie**", "**Orthologie**" verglichen werden – wie gehen wir vor?

PCR-Varianten:

→ Die **Emulsions PCR** ist eine wichtige Technik. Es werden Adapter auf beiden Seiten des Gens ligiert. Dann wird die DNA denaturiert und man gibt eine präzise verdünnte Lösung der DNA auf eine Glasoberfläche hinzu, wo sich die einzelsträngigen DNA-Stränge entweder miteinander verbinden, oder sie verbinden sich mit kleinen Oligonukleotiden auf der Glasoberfläche, wodurch die DNA fest mit der Glasoberfläche verbunden wird (= **Immobilisierung**). Die Stärke dieser Bindungen ist relativ gering.

Entweder binden sie an zwei **verschiedene Oligonukleotide** (je nach Temperatur) und bilden eine Brücke oder es wird keine Brücke erstellt. Durch Aufschmelzen kann man das Loslösen gezielt steuern. Somit geschieht die PCR durch kleine Schritte durch diese Anheftung der Oligonukleotide, wodurch die Nukleotidkette ständig weiter wächst. Am Ende entstehen aus den ursprünglichen Einzelsträngen mehrere **Cluster**.

Eine weitere wichtige **Weiterentwicklungen** ist die Detektion der Sequenzierung. **Sequenzierung durch Synthese** ist ein wichtiges Thema. Gilt nur für DNA, da Helix Struktur.

SEQUENCING:

Sanger Sequencing: → Die Begrenzung auf 800 bp leitet sich dadurch her, dass der relative Unterschied (prozentuell) zwischen den Molekülen zu klein wird und die Banden miteinander verschwimmen.

Die klassische **Sanger Sequencing** war zudem ein Problem, da nicht jedes Fragment von E. Coli aufgenommen werden kann – manche Inserts sind für den Organismus toxisch.

Abhilfe schuf hier **Next generation sequencing**:
→ Hier wird die Klonierung in *E. coli* ausgelassen und eine PCR durchgeführt. Bis auf Enzyme, Polymerasen, Primer und einige andere Faktoren ist die PCR eine rein künstliche im Reaktionsglas durchführbaren Reaktion. Somit gibt es keine Toxizität mehr in bezug auf die Verwendung lebendiger Systeme (bei dem **Klonierungsschritt**).

PacBio: arbeitet mit einzelnen **Molekülen**. Die Polymerase selbst wird hierbei fixiert. Nachteil der Methode: PacBio ist **extrem teuer**. Die Fehlerrate ist rein zufällig, somit könnte man mit **mehrfacher Sequenzierung** die Fehler statistisch ausschließen.

Sequence assembly:

→ Beim **Shotgun Assembly** gilt: je mehr Basen miteinander überlappen, desto **wahrscheinlicher** ist eine Zusammengehörigkeit.
→ Man muss bei der Sequenzierung allgemein davon ausgehen, dass eine **Fehlerquote** besteht. Der **Qualitätswert** gibt wiederum an, wie **sicher** eine Sequenzanalyse ist. Beispielsweise ist ein Wert von ~30 relativ gut, darunter werden die Werte immer ungenauer. Dies ist bei der Sequenzierung durch Deutlichkeit der Fluoreszenz bzw. Klarheit des Signals erkennbar. Beim **Pyrosequencing 454** ist es der zeitliche Verlauf, der die Fehlerwahrscheinlichkeit wiedergibt.

De novo assembly

Die reads werden fragmentiert geliefert. Mit diesen einzelnen Teilen müssen wir die ursprüngliche Reihenfolge wiederherstellen. Wenn die **Überlappungen** übereinstimmen, können wir die Kette aus bp wieder herstellen.

Legt man die **Kettenfragmente** übereinander, so gibt es bei den einzelnen bp Übereinstimmungen, aber auch Unterschiedlichkeiten. Wenn an einer Stelle zu 50% G und 50% A angezeigt wird, kann es sich um einen single nucleotide polymorphism handeln – also zwei Allele. Falls z.B alle reads an einer Stelle ein C haben und nur ein einziges A vorkommt, wird es sich je nach Häufigkeit wohl um eine **Sequenzierfehler** handeln.

Assemblies sind daher genauer als einzelne reads, da man einen Abgleich auf statistischer Basis hat. Man ist also genauer.

Probleme:
Sequenzwiederholungen; Gleiche Abschnitte können an vielen Positionen richtig eingesetzt werden. Nur einmal vorkommende Sequenzen dagegen nicht. Daher führen repeats zu fundamentalen Problemen, wofür es leider kaum eine Lösung gibt.

Außerdem kommt es schnell zu Begrenzungen durch die **Hardware der Computer**. Die **Assemblies** benötigen **Arbeitsspeicher** von mehreren Terabytes.

Fastqc: reads quality statistics.

Mapping: Wir nehmen ein Genom aus öffentlichen Datenbanken und verwenden es als **Referenz** zu einem unbekannten Genom. So kann man mithilfe einer Vorlage schon bekannte Abschnitte erkennen.

Paired-End Sequencing:

Paired-end sequencing allows users to sequence both ends of a fragment and generate high-quality, alignable sequence data. Paired-end sequencing facilitates detection of genomic rearrangements and repetitive sequence elements, as well as gene fusions and novel transcripts. Since paired-end reads are more likely to align to a reference, the quality of the entire data set improves. All Illumina next-generation sequencing (**NGS**) systems are capable of **paired-end sequencing**.

Important terms:

Sequence read	Raw sequencing result for one DNA target ("**insert**")
Paired end reads	Two reads starting on both ends of the same DNA PCR target ("**insert**")
Mate pair reads	Two reads having a known distance in the original DNA sequence
Base quality	Numeric value indicating the **quality** of the sequencing signal
Filtering	Discarding sequence reads of overall low base quality; hier erhält man also nur reads bestimmter Qualität
Trimming	Shortening sequence end(s) to **improve overall base quality of a read**. Durch **trimming** (=Zurechtstutzen) werden Reste, die z.B. eine kleinere **Qualität als 20** haben, einfach weggeschnitten und nicht weiter verarbeitet.
Assembly	Grouping unambiguously overlapping reads into contiguous sequence
Contig	Contiguous DNA sequence derived from assembly; may contain ambiguous/unknown bases
Scaffolding	**Grouping contigs** into longer sequences using e.g. mate pair sequences
Scaffold	DNA sequence consisting of contigs in a defined order; may include gaps of estimated length
Finishing	Closing gaps, resolving ambiguities/unknown bases and linking contigs into final genomic sequences

Polymorphe Hotspots: diese Regionen sind **hochvariabel**, vor allem bei Mikroorganismen.

Prediction of non-coding genes / Genvorhersage:

Bakterielle Gene starten meistens mit **ATG**; es gibt jedoch bis zu fünf verschiedene **Startcodons**. ATG ist das häufigste, aber nicht das einzige. In E. coli sieht die Situation wie folgt aus:

Start codons in E.coli

ATG:	**3542 (82.6%)**
GTG:	**612 (14.3%)**
TTG:	**130 (3%)**
ATT:	**1**
CTG:	**1?**

Gegeben sei eine **DNA Sequenz** von *E. coli*. Es sind nur **A**, **T**, **G** und **C** zu sehen. Eine Unsicherheit bei der Sequenzierung wird normalerweise mit anderen Buchstaben dargestellt. Ein Beispiel wäre "**N**" für "Unbekannt" oder "nicht sequenziert". Für jede andere Gruppe von Kombinationen gibt es einen **eigenen Buchstaben**. "**R**" wäre für A oder G. "Y" wäre für C oder T etc. Es gibt einige weitere Buchstaben.

Nicht jeder ORF ist automatisch ein Bereich, der für Proteine kodiert. Grundsätzlich ist jeder ORF ab 600 Nukleotide ein potentielles Gen, das für ein Protein codiert. Darunter liegende ORF's bis ca. 600 Nukleotide können oft nur eher zufällig sein. Eine **Filterfunktion** für alle unter 600 Nukleotide liegenden Abschnitte ist allerdings **nicht sinnvoll**, da die Information dennoch wichtig sein können.

Wie geht man bei einer Sequenz mit der **Doppelsträngigkeit** um (DNA wird ja in Form eines Doppelstranges zu finden sein)? Günstig wäre ein **Startpunkt im Genom**, vor allem bei **bakteriellen zirkulären Genomen**. Man will keine Gene in der Mitte spalten, also muss man sich einen **Ort** suchen, der günstiger ist – ein Beispiel wäre die Zone, an der die Replikation (***origin of replication***) der DNA normalerweise beginnt. Sequenzierungen werden allerdings an verschiedenen Punkten begonnen.

Wir müssen nun entscheiden, **welchen Strang** wir nehmen. Auf beiden Strängen können gleichermaßen Gene kodiert werden. Es gibt **keinen** "**wichtigeren Strang**" für die Zelle. Bei Gen-Darstellungen wird der Ort der Gene mit Pfeilen dargestellt. Je nach Richtung wird der Strang dargestellt.

5' zu 3' wird allerdings von der Polymerase bevorzugt, darum nehmen auch wir diese Richtung. Kennen wir schon bekannte Sequenzen, so entscheiden wir uns für den selben Strang. Grundsätzlich hat man sich zufällig für den einen Strang bei der ersten Sequenzierung entschieden. Daran hat man sich zur Verhinderung von noch mehr Verwirrung dann auch gehalten.

Zur Genvorhersage:

Wir wollen von der Sequenz zu einer Liste von Genen und in weiterer Folge zu Funkionen kommen.

Die **rRNAs** gehören zu den **nicht codierenden Genen**, werden aber oft als **evolutionärer Marker** verwendet. Sie sind hochkonservierte DNA Sequenzen. Mithilfe von BLAST kann man die Regionen mit bereits bekannten DNA Abschnitten vergleichen.

Man kann sowohl **5S rRNA** (~120nt), 16S rRNA (~150nt) als auch 23S rRNA (~2900nt) verwenden. Allerdings werden die 5S rRNAs von den Datenbanken nicht unterstützt, weil sie sich aufgrund der kurzen Länge nicht eignet.

RNAmmer kann **alle drei rRNAs** behandeln. Dabei haben wir positionspezifische Konservierungsmodelle für diese rRNAs. BLAST hat diese nur für 2 von 3. Wenn man weiß, in welchen Teilbereichen man hochkonservierte DNA Abschnitte findet, ist es damit relativ leicht die DNA zu interpretieren und Gene vorherzusagen.

Ausnahmen und Besonderheiten: Selbstsplicende Introns, die sich aus dem Transkript von selbst entfernen. Diese sind momentan noch relativ unerforscht und unbeobachtet. In den nächsten Jahren wird hier noch intensiv geforscht werden.

Es gibt auch noch kürzere nicht codierende RNA Abschnitte: z.B tRNAs. Sie sind kürzer (74-95 nt) und in der Sequenz noch viel variabler. Mit BLAST oder Sequenzkonservierung kommen wir hier daher nicht mehr weiter. Dies ist daher der erste Gentyp, wo wir überlegen müssen, was wir neben der Sequenz noch verwenden können, um eine Unterscheidung treffen zu können. Wir behandeln daher die Sekundärstruktur.

Wir benutzen mit tRNAscan-SE (Programm) sowohl die Sequenz als auch die Sekundärstruktur der tRNA.

Bis vor 5 Jahren dachte man, dass man die nicht-codierenden Abschnitte mit rRNA's und tRNA's abdecken könne. Allerdings sind Bakteriengenome voller nicht-codierender Abschnitte. Wir sind grade mal am Anfang diese Abschnitte zu verstehen. Viele haben z.B. regulatorische Funktion.

RFAM ist eine Datenbank die mittlerweile mehr als 2000 Familien von nicht codierenden RNA's enthält. Jede wird mit Sequenz und Struktur beschrieben. Von den meisten haben wir allerdings noch keine Ahnung, was sie tun.

Ein Beispiel: **RnaseP** (**Ribozym**) wird wie ein Wiki dargestellt. Zusammenarbeit mit Wikipedia. Eine der aktuellsten Seiten, was den Wissensstand angeht.

Praktische Realisierung der Gen-Vorhersage:

Zuerst suchen wir nach allen **ORFs**, die in der Genomsequenz vorliegen **können**. Davon nehmen wir die mit **ausreichender Länge** heraus. Dadurch lernen wir die Codon-Usage des Organismus kennen. Wir entwickeln eine für diesen Organismus gültige Häufigkeit von gewissen Codons, um in weiterer Folge Gene vorhersagen zu können.

Dieses **statistische Verfahren** wird verwendet um die Starts abzuleiten und dann ein **erstes Modell für die ribosomalen Bindestellen zu erstellen**.

Somit wird die **Statistik** der Nukleotide und der Codons und die Gestalt der ribosomalen Bindestellen hierdurch kennengelernt. Man nennt dies "**self-training**", wodurch die Software quasi selbst Algorhihmen entwickelt.

Mit diesen **statistischen Modellen** kann man dann aus allen **ORFs** erschließen, welche für weitere Analysen behalten werden, und welche nicht.

Was für **Probleme** ergeben sich aber? Lange Abschnitte werden gleich beurteilt wie kurze, was zu falschen Ergebnissen führen kann. Außerdem gibt es eine evolutionäre Kritik: Es gibt Gene, die horizontal weitergegeben wurden. Manche Plasmide werden übergeben oder ausgetauscht.

Die **Codon-Statistiken** stimmen dann natürlich überhaupt nicht überein. Es braucht einige Zeit, bis sich die Gene adaptieren. Vor allem bei frischen Gentransfers kann man keine Annahme machen. Man muss daher von Vornherein zwei Modelle schaffen. Einerseits die typische Nukleotidzusammensetzung und andererseits die atypische (Sonderfälle im Genom).

Zusammenfassend:

Ein Modus mit reiner Sequenzanalyse (**rRNA's**) und ein Modus mit Kombination von Sequenz und Struktur (→ **RFAM**). Dies betrifft **nicht-codierende RNA Gene**.

Es gibt insgesamt **vier verschiedene Arten** von **offenen Leserahmen**:

0 – **Translation** zwischen **Stopcodons**
1 – **Translation** zwischen **Start**- und **Stopcodon**
2 – **Nukleotidsequenz** zwischen **Stopcodons**
3 – **Nukleotidsequenz** zwischen **Start**- und **Stopcodon**

Amino Acid Substitution Matrices

High conservation expected:
- BLOSUM80
- PAM120

Low conservation expected:
- BLOSUM50
- PAM250

BLOSUM62, for instance, has costs of **gap open: -10**; **gap extent: -2**.

Organismus	**GC content des Genoms**
E. coli (K12)	**Etwa 51% GC Anteile**

Tools for Gene prediction:
Metagene, GeneMark, ORF-Finder, FragGeneScan, fgenesB, GLIMMER, BLAST
Sequence assembly:
Phrap, Forge, Arachne, JAZZ, Celera, Velvet, Newbler, SOAPdenovo, EULER, ORFome assembly, IDBA-UD
Taxonomic binning:
MetaBin, MEGAN, WebCARMA, PhyloPythia, TETRA, NBC, TACOA
RNA gene prediction:
tRNAscan-SE

Glossar	Funktion
CMs:	**Covariance Models**
Global alignment:	**all residues aligned (Needleman-Wunsch algorithm)**
Local alignment:	**most similar regions aligned (Smith-Waterman algorithm)**
ORF Finder:	**Allows us to find ORFS → http://www.ncbi.nlm.nih.gov/gorf/gorf.html**
CRITICA:	**Coding Region Identification Tool Invoking Comparative Analysis**
RFAM:	**Datenbank für nicht-codierende RNAs. Diese Datenbank inkludiert fast 2700 Familien.**
Homologie:	**Ähnlichkeit kommt vom gemeinsamen Vorfahren.**
metagenomics:	**culture independent technology to study microbes inhabiting different environments**
psiBLAST:	**position specific interacted BLAST**
PRODOM:	**sekundäre Datenbank, die auf psiBLAST beruht**
RNAmmer:	**Useul for position-specific sequence conservation models**

Slogans:

→ Tipp für die Prüfung: wenn keine Information gegeben, wie teuer ein Experiment sein kann, dann eine **Annahme treffen** und unbedingt anmerken. Wir haben keinen Vergleich, da neue Spezies → daher streichen wir "**comparative assembly**" und "**mapping of genomes**". Es bleibt uns nur noch die "**de-novo assembly**".

→ Der **Boden** ist eine **Grenzschicht** zwischen Lebensräumen/Zuständen (flüssig-fest-gasförmig).

→ **Microbes** are **ubiquitous** in **every habitat on earth**: *Water*, *Soil*, *Deep* in the *earth's crust*, *organic matter*, *polar ice shelf*, *acidic hot springs*, *radioactive waste*, with and within eukaryotes.

→ **Dotplots** erlauben es einem nach längeren Übereinstimmungen zwischen zwei Nukleotidsequenzen zu suchen.

→ **In-vivo cloning** thanks to **bacterial artificial chromosomes** (**BACs**). **sstDNA** can be annealed to an **excess of DNA Capture Beads**.

→ **Proteindomänen** unterstreichen den **modularen Aufbau eines Proteins**.

→ Eine Fehleingabe der Gene, die es eigentlich gar nicht gibt, führt in der Datenbank zu **falschen Proteinfamilien**.

→ Most **annotation** is computated automatically these days, at the least initially.

→ **Fehlerraten** können oft durch **mehrere reads** neutralisiert werden.

→ **Gelöschte Daten** können nicht mehr verwendet werden, darum werden wie oft in der Forschung **mehr Informationen** geliefert, als nötig - die **Anpassung** erfolgt anhand der spezifischen Fragestellung.

→ **Optical mapping** requires **fluorescent staining**.

→ reads should align completely, without insertions
→ Similarity may imply similar functionality.
→ **Poorly adapted codons** at the N terminus of genes may slow down ribosomal elongation during initiation.
→ Nicht jedes bakterielle Gen hat direkt einen eigenen Promotor. Insbesondere geclusterte Gene (**Operons**) werden oft nur durch **einen Promotor** vorab der Sequenz dirigiert.
→ **repeats** should have the same coverage as non-repeat
→ **ClustalW** creates a **guide tree**.
→ **GenBank** is highly redundant.
→ **Heuristic methods** helped apply simplifying assumptions to the problem of searching for sequences.
→ Ein "**evolutionärer Aspekt**" ist zum Beispiel dann gegeben, wenn bestimmte Aminosäuren unter **Selektionsdruck** stehen, andere hingegen nicht.
→ **ORFs** sind immer gegeben, erst die Länge macht den jeweiligen ORF interessant.
→ **Multiple sequence alignments** werden bei mehr als zwei Sequenzen benötigt.
→ mRNAs in Bakterien besitzen die **Shine-Dalgarno-Sequenz**, die man auch heranziehen können zur **Genomanalyse**
→ Ein Stopcodon ist nicht in allen Spezies für die Stopfunktion da.
→ **Pairwise sequence alignment** allows us to **look back billions of years ago**.
→ **RBS positional frequency pattern** can be detected by **GeneMarkS**.
→ Because the 5S rRNA, 16S rRNA and 23S rRNA is often very highly conserved on the DNA level, this may allow identification through **BLAST**.
→ **BLAST** gibt **lokale Alignments** aus (das L in BLAST steht für Local).