(1) Bioinformatics for Biomedical Engineers VO (Mach, (TU Wien), 166.229

Die Prüfung "**Angewandte Bioinformatik**" bzw. "**Bioinformatics for Biomedical Engineers**" an der **TU Wien** besteht aus **10 Fragen** .

Ein (mögliches) **Passwort** besteht in der Regel aus der **LVA-Nr** (mit Punkt in der Mitte), also: **166.229**

Inhalt der Lehrveranstaltung:

- Einführung in die Bioinformatik
- Grundlagen der **Statistik**
- Werkzeuge & Algorithmen zur Suche ähnlicher Sequenzen
- Sequenzabhängiges Design (**PCR**, Melt....)
- **Datenbanken**
- Modelle molekularer Evolution
- **Multiples Alignment**
- **Genomics**
- **Protein**: Sequenz, Struktur, Funktion
- Klassifikation, Identifikation & Phylogenie
- Demonstrationsbeispiel Sequenzanalyse Datenbankabfrage
- Demonstrationsbeispiel Musteranalyse (Gelauswertung)
- Smith-Waterman Algorithmus
- Genomic & in silico Analyse des Genoms
- Genomprojekte Aufbau
- Arbeitstechniken und Datenverarbeitung
- Identifikation von Organismen
- Molekulare Taxonomie Datenbanken
- Aufbau und Abfrage in Datenbanken
- Das **WWW** als Informationsquelle für Gentechniker und Biochemiker

Hilfreiches Videos bezüglich **Bioinformatik**:
http://www.youtube.com/watch?v=OnpoJ2aysgY

Mnemo-Sachen:

Die **Edit-Distance** erlaubt **3 Operationen**:
(1) **Insert** (2) **Delete** (3) **Replace**

Prüfungsfragen, Beispielfragen und Trainingsfragen:

:: What is the **difference** between a "**motive search**" and a "**pattern search**" in a peptide? Worauf beruht die Motivsuche in Proteinen, Bsp für ein Motiv, wie werden **Motive** in DB abgelegt, welche Nachteile existieren? Was gehört zu **Sekundär-Struktur** eines Proteins (Coiled Coil, Helix Turn Helix)

Motive search:
→ **Proteins** are composed of **functional units** (domains, motives).
→ Such **functional units** are preserved temporally and **across species borders**.
→ **Exon-splicing theory**: Exons are protein fold domains which have different additional functions.
→ Databases with **collections of known motives**: = **PROSITE**.
→ Every domain is filed as an easy pattern and linked with additional information of known proteins bearing such a motive.

Example:
ATP/GTP Binding motive in **PROSITE**: [AG]-X(4)-G-K-[ST] → http://prosite.expasy.org/cgi-bin/prosite/prosite-search-ful?SEARCH=-G-K-&makeWild=on, → http://prosite.expasy.org/PDOC00017

Drawbacks:
(1) • No **new variations** are found.
(2) • Simple motives lead to **randomized detection** and **false positive identification**.

Profile search – pattern (motive) search:
→ Performing a pattern search one goes for motives, derived from known proteins, (new variations are overlooked).
→ Performing a profile search it is searched for similarities to known domains.
→ **Profiles** are **tables** of **amino acid frequency** for every position of a motive.
→ Profiles are derived from multiple alignment analyses of preserved domains from members of a protein family.

:: Welche ist die **richtige Reihenfolge**? (Es folgten mehrere Reihenfolgen der Klassifizierung in Kingdom, Species, Order usw.) **3 Domains of Life?** Worauf beruhen sie? [**DONE**]

Reich → Stamm → Klasse → Ordnung → Familie → Gattung → Art
Kingdom → Phylum → Class → Order → Family → Genus

:: **ESTs** sind... (es folgte eine **Auflistung von Fakten**, von denen man die richtigen ankreuzen musste, darunter auch wofür die Abkürzung stehen kann) Anwendungsbeispiele der ESTs nennen. [**DONE**]

ESTs is an abbreviation for "**Expressed Sequence Tags** ". An **expressed sequence tag** (**EST**) is a short sub-sequence of a **cDNA sequence**. They derive from a mRNA sequence, but only from the exons. There are more than 50 million ESTs known since about 2010.

One **advantage** of ESTs is that they are not very expensive, and they can be used to discover new genes.

Applications for **ESTs:**

(1) **Genome Landmarks**: Da ESTs meist Sequenzen repräsentieren, die in einem Genom nur einmal vorkommen, können sie als Orientierungspunkte beim Zusammenbau der im Rahmen einer Genomsequenzierung anfallenden Sequenzdaten dienen.
Dies spielte z.B. bei der Entschlüsselung des menschlichen Genoms eine entscheidende Rolle.

(2) **Expressionsanalysen**: Aus dem Auftreten und der Häufigkeit von ESTs lassen sich grobe Informationen zur Expression der betreffenden Gene gewinnen. Dieses Verfahren wird teilweise in Anlehnung an den Northern Blot als Virtual Northern Blot bezeichnet.

(3) **Identifikation unbekannter Gene oder Finden seltener Gene**: Durch den Vergleich von ESTs mit den Sequenzen bekannter Gene können verwandte Gene im gleichen oder in anderen Organismen identifiziert werden. Die **ESTs** können dann als **Gensonde** dienen, um die Gesamtsequenz des betreffenden Gens z.B. in einer cDNA-Bank oder mittels **RACE-PCR** zu gewinnen.

(4) **ESTs** can be mapped to **specific chromosome locations** using physical mapping techniques, such as radiation hybrid mapping, Happy mapping or FISH.

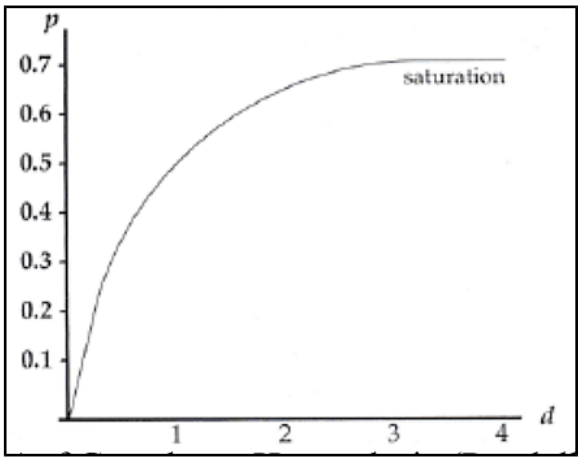
ESTs kann man über **dbEST** → http://www.ncbi.nlm.nih.gov/dbEST/ abfragen.

ESTs sind kurze (200-500bp) cDNA-Stücke, die über RT – PCR aus der mRNA von Zellen unterschiedlicher Gewebe oder von Zellkulturen oder Organismen unterschiedlicher Kultivierungen stammen.

Es werden **EST Banken** angelegt und die klonierten Fragmente die aus codierenden Regionen (**Exons**) stammen und nur einmal im Genom vorkommen (keine Repeats).

:: What is the **difference** between p(observed) and d(estimated) **genetic distances**? Draw a scatter plot showing how these values correspond to each other. Explain why it is worthwhile to use both measures. [DONE]

p(observed): **Anzahl** an**,** zur Zeit, abzählbaren **Punktmutationen**
d(estimated): **Anzahl** an **Punktmutationen** gesamt (mit hin- und Rückmutation)



Auf Grund von **Homoplasie** (Paralell- oder Rückmutationen) unterschreitet die beobachtete Anzahl (p) deutlich die geschätzte Zahl (d).

observed: Daran kann man den Unterschied bzw. die Ähnlichkeit feststellen.
estimated: (inkl. ***Clock Theory***) An der Gesamtzahl kann man feststellen wie lange es her ist, dass sich die beiden Unterscheiden, bzw. welche Zwischenstufe früher da gewesen sein muss.

:: Erklären Sie die **Funktionsweise** von DNA-Chip **Microarrays**, welche Aussagen können durch den Einsatz von DNA-Chip Microarrays erfolgen, nennen sie die **wichtigsten Anwendungen** (mindestens 5 Anwendungsbeispiele der **DNA Chips**)? [DONE]

• Eine **Vielzahl** von cDNA Fragmenten oder Oligonukleotiden (derzeit 100.000 bald 1.000.000 und mehr) werden auf **Glas Objektträger** in einer Reihe gebunden (**Chip**).
• Eine **fluoreszenzmarkierte RNA Probe** wird mit dem Chip **hybridisiert**.
• Die relative Menge an **exprimierter mRNA** wird über **die Messung der Fluoreszenz** des jeweiligen Auftragepunkts gemessen.

Anwendungen:

→ Suche nach Stoffwechselwegen
→ **Suche** nach **genetisch verursachten Krankheiten** / **Krebsmuster**
→ **Sequenzieren** von DNA Molekülen

:: **SNPs** sind... (es folgte eine **Auflistung** von Fakten, von denen man die richtigen ankreuzen musste, darunter auch wofür die Abkürzung stehen kann), Anwendungsbeispiele der SNPs aka "wofür werden SNPs verwendet?" [DONE]

SNPs: = Single-nucleotide polymorphisms:

SNPs sind **DNA-Punktmutationen** (Basenaustausche, Insertionen oder Deletionen) die als **die häufigste Variation in Genomen** auftreten. Verschiedene Mitglieder der selben Spezies weisen hier Unterschiede in der Basenzusammensetzungen auf.

SNPs existieren an definierten Positionen im Genom, den → **STSs** („***Sequence Tagged Sites***").

Nehmen wir **zwei DNA Sequenzen**, zwei Allele:
AAGC**C**TA
AAGC**T**TA
Hier sieht man einen Unterschied an einer Position – das ist somit ein SNP.

Die Frequenz von SNPs ist messbar und liegt bei +/- 1000 bp.

Die **wesentlichen Anwendungsgebiete** von SNPs sind:
„**Mapping**" von Genen
Definieren von **Populationsstrukturen**
Funktionelle Studien von **Genen**
Untersuchung des Zusammenhangs von Sequenzvariationen und erblichen Phenotypen
juristische bzw. **forensische Anwendungen**

Die Detektion von SNPs ist möglich durch direktes Sequenzieren oder z.B. DNA Chips (für den Nachweis einer Vielzahl von SNPs in einem Ansatz)
Vorteile von SNPs gegenüber anderen **genetischen Markern**:
• Einfache und unaufwendige Nachweistechniken
• Stabile Vererbung nach Mendelschen Gesetzen
• Geringe Spontanmutationen

Link: → http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

:: Was besagt das "**LAW OF PARSIMONY**"? Was ist **PARSIMONY**, was ist **MINIMUM EVOLUTION**? What is the law of parsimony and parsimony (**minimum evolution**). Explain how it is applied in phylogentic reconstruction. What is "Tree Reconstruction"? [**DONE**]

Parsimony = Sparsamkeit

Das **Law of Parsimony** besagt das die **einfachste** der plausiblen (gleichrichtigen) Theorien vorzuziehen ist.

PASRSIMONY (**MINIMUM EVOLUTION**): Der **phylogenetische Baum**, bei dem die Individuen die **wenigsten Änderungen** durchmachen soll, wird bevorzugt.

Das "**Law of Parsimony**" wird auch **Occam's razor** genannt, nach "**William of Occam**".

In der **Wissenschaft** allgemein lässt sich das *law of parsimony* auch so formulieren:

> *The principle that **assumptions** introduced to explain a thing **must not be multiplied beyond necessity**, and **hence the simplest of several hypotheses is always the best** in **accounting for unexplained facts**.*

Maximum Likelihood: Wir suchen den Baum der uns mit der **größten Wahrscheinlichkeit** unsere Daten generiert haben könnte.

:: Berechnen der **Laplace-Wahrscheinlichkeit** und des **Erwartungswertes**. Welche **Wahrscheinlichkeit** dass 1 und 2 bei einem Würfelexperiment (Würfelseiten bezeichnet mit 1, 2, 3, 4, 5, 6, 7, 8) zutrifft? Was ist der zugehörige **Erwartungswert** bei 10 Würfen? Wann spricht man von einem **Laplace-Experiment**? Wie verändert sich das ganze wenn wir einen sechseitigen Würfel verwenden? Geben Sie die Formel für die "**Wahrscheinlichkeit**" und den "**Erwartungswert**" an.

Definieren wir zuerst die **Laplace-Wahrscheinlichkeit**. Man spricht von einem **Laplace-Experiment** (auch "Gleichwahrscheinlichkeit" genannt) wenn **jedes Einzelereignis mit der selben Wahrscheinlichkeit auftreten kann**, wie zum Beispiel bei einem **Münzwurf** wo **Kopf** oder **Zahl** mit einer Wahrscheinlichkeit von **je 0.5** auftreten können.

Die **Anzahl der möglichen Fälle** wird mit **W** bezeichnet.

Haben wir einen **sechseitigen** Würfel, so ist die Menge der möglichen Fälle wie folgt:
W = **{ 1; 2; 3; 4; 5; 6}**, hier enthält also **n = 6 gleichwahrscheinliche Elementarereignisse**. In dem Fall hier 1 / n, also 1 / 6.
Bein einem **achtseitigen Würfel** gibt es natürlich mehr Elemente der Menge W, und somit für jedes Einzelereignis eine **Chance von 1/8**.
Das nun 2 Ereignisse davon eintreten ist 1/8 + 1/8, also 2/8 → 1 / 4, also **0.25**.

Die Anzahl der günstigen Fälle bezeichnet man mit **E**, für "**Ereignis**" (oder "Event"). Für ungerade Zahlen eines **Sechsseiters** gilt:
E = {**1**} {**3**} {**5**}

Umgekehrt gilt natürlich das gleiche für die geraden Zahlen, also **2**, **4** und **6**.

Das **Ereignis E** "ungerade Augenzahl" ist als Vereinigung von g = **3 Elementarereignisse** darstellbar.

Die **Formel** für den **Erwartungswert** ist allgemein:

E = **Anzahl** an **Möglichkeiten W**
E = Σ ni x Pi (E)
E = 100 * 1/8 = 12,5 [für einen Würfel mit 8 Seitenflächen)
P(E) = Σ günstigen Fälle (E) / Σ aller möglichen Fälle (Ω)
P(E)=2/8=1/4

Wie geht man **allgemein** bei so einer **Fragestellung zu Laplace** vor?

Man überlegt sich **zuerst wieviele Möglichkeiten** es gibt. Im obigen Beispiel mit dem **Achtseiter** wären es eben **8 Möglichkeiten**, die alle gleich wahrscheinlich sind. Bei einem Sechsseiter sind es eben **6 Möglichkeiten**.

Sehen wir uns ein weiteres Beispiel an. Zwei Buchstaben werden aus dem Wort "**LASSO**" zufällig und **ohne Zurücklegen** ausgewählt.
Wie **groß** ist die Wahrscheinlichkeit dafür, dass **zwei Konsonanten** gewählt werden?

LASSO hat **5 Buchstaben**, darunter 3 Konsonanten und 2 Vokale. Das **zwei Konsonanten** gewählt werden ist 1 x (3/5), sowie 1x (2 / 4), (3 * 2 / 5 * 4), also 6 / 20 → 0.3.

→ http://www.youtube.com/watch?v=sqAnRdeHhwE
→ http://www.youtube.com/watch?v=snJf9cxveyE
→ http://www.serlo.org/math/exercises/topics/show/Stochastik/Wahrscheinlichkeit/Laplace-Wahrscheinlichkeiten
→ http://mathenexus.zum.de/formelsammlungen/stochastik/S301LaplaceExperiment+Wahrscheinlichkeit.htm

:: Nennen Sie **3 wichtige Bioinformatikdatenbanken**. Nennen Sie die **drei wesentlichen Kategorien** von **bioinformatischen Datenbanken** und nennen Sie ein **Beispiel** der darin enthaltenen Informationen (**KEINE Bioinformatik Server nenen**!) Which types of **queries** for **sequence databases** do you know (4 Examples)? What **types of queries** would you recommend? Why? Was ist die Rolle von **KEGG** in der **bioinformatischen Metagenomanalyse**? Worin unterscheidet sich **KEGG** von anderen Datenbanken? Was wird in Bioinformatik-Datenbanken gespeichert?

Unterschieden werden kann zwischen **primären** und **sekundären Datenbanken**.

(1) **Protein Data Bank** (**PDB**)
(2) **PIR** ("*Protein Information Resource*")
(3) Swiss-Prot / **UniProt**
(4) **GenBank**
(5) **Flybase** (Drosophila)
(6) **Glimmer** → http://www.cbcb.umd.edu/software/glimmer/
(7) **KEGG –** mittels **BRITE** zeigt es **biomolekule Wechselwirkungen an.** (**KEGG** = ***Kyoto Encyclopedia of Genes and Genomes***)
(8) **EBI**
(9) **NCBI**

→ **KEGG** consists of **16 different database**. **KEGG** has been widely used as **a reference knowledge base** for biological interpretation of large-scale datasets generated by sequencing and other high-throughput experimental technologies. **5 KEGG databases** would be: **KEGG PATHWAY, KEGG BRITE, KEGG MODULE, KEGG DISEASE, KEGG DRUG**.

→ **Glimmer** sequentially scans nucleotide sequences for **particular kmers** (e.g. the **5mer ATGGC**) and estimates the probability of that pattern occurring in a real gene. The **statistical** model of a gene is then used to analyze the complete set of unknown genomic DNA. The **ORFs** that are analyzed by Glimmer must exceed some **minimum length** (e.g. 99 base pairs).

Glimmer uses **a hidden Markov model** (**HMM**) **approach**. HMMs are statistical models of the patterns of nucleotides comprising a gene. The **HMM** includes observed states (e.g. nucleotide sequence including a start or stop codon) and hidden states (genes in DNA).

Queries: **Glimmer** scans for particular kmers.
→ http://en.wikipedia.org/wiki/List_of_biological_databases

:: Eine **Substitutionsmatrix** für den Vergleich von Aminosäurensequenzen beruht auf dem Prinzip der/des: [**DONE**]

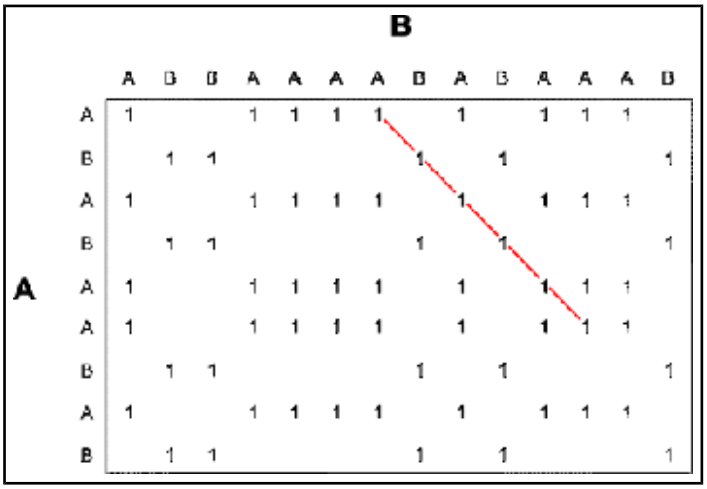
a) Buchstabencodes der AS
b) relativen Isotopenzusammensetzung von AS
c) Sequenzabfolge der AS
d) chemischer Ähnlichkeit von AS (← richtige Antwort)

:: **Zeichnen** Sie einen **Dotplot** für **A**:= **DDCCDDCD**, **B**:= **CDDCDDC** codieren ai = bj = 2 und ai != bj = 0. Markieren Sie **den längsten Teilstring!** Was ist der Einsattzweck von Dotplots? Zeichnen Sie einen Dotplot mit den Zeichenketten A:=CCDDCD und B:=CDDCDDC und codieren Sie mit ai entspricht bi = 2, sowie ai entspricht nicht bi = 0. Markieren Sie den längsten Teilstring. **[DONE]**

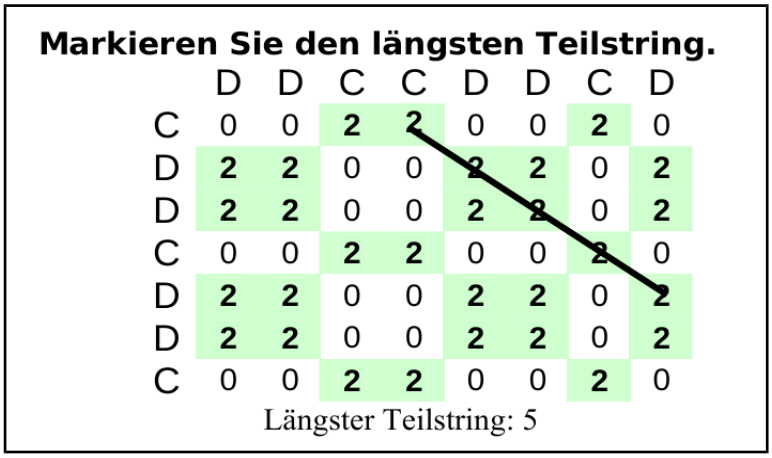
Ein **Dotplot** ist eine **graphische Methode**, die es uns erlaubt **2 biologische Sequenzen** zu vergleichen und nach Regionen mit grosser Ähnlichkeit einzuteilen. Die Angabe oben heisst das für ein match der Wert von 2 verteilt wird, und für ein nicht-match ein Wert von 0.

Filterung kann die Erkennung von **Übereinstimmungen** erleichtern.

Auch **Anordnungen** von Genen in Genomen zwischen Organismen können verglichen werden.

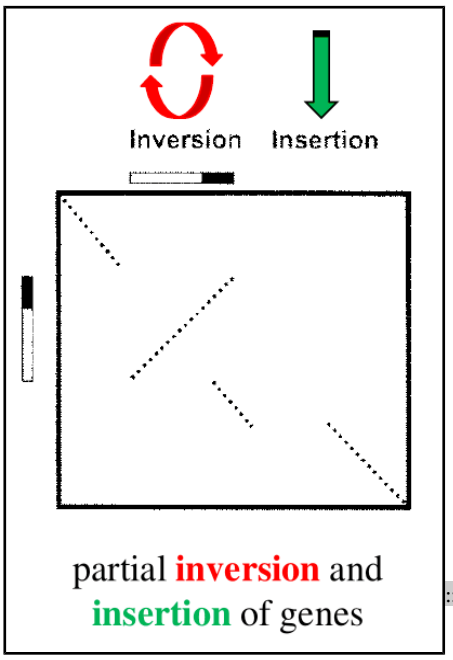


Hier die Lösung zum Problem:



→ http://www.youtube.com/watch?v=YWV056ddOB4

We can identify an **inversion** in a dotplot by a 180° degree shift.



:: Explain **global** vs. **local** alignment by a **simple diagramm**. What is an **Alignment**? Welche **2**

Algorithmen ermöglichen die Berechnung der Distanzen bzw. Ähnlichkeiten und die Ableitung ihres Alignments zwischen 2 **DNA Sequenzen ungleicher Länge**?

Ein **Alignment** ist die **Ausrichtung** bzw. die **Arrangierung** von identischen oder ähnlichen Buchstaben **zwischen zwei** oder **mehreren Sequenzen**.

OTTO ist ident zu **OTTO** aber nicht völlig ident zu **TOTO**.

Global alignments:
→ They attempt to **align every residue in every sequence**.

They are most useful when the sequences in the query set are similar and of roughly equal size.

Local Alignments:
→ local alignments identify regions of similarity within long sequences that are often widely divergent overall.

Local alignments are often preferable, but can be **more difficult to calculate** because of the additional challenge of identifying the regions of similarity.

A general global alignment technique is the **Needleman–Wunsch algorithm**, which is based on **dynamic programming**.
Local alignments are more useful for dissimilar sequences.

The **Smith–Waterman algorithm** is a general local alignment method, and it is also based on **dynamic programming**.

:: Welche **evolutionäre Modelle** gibt es? Welche **Parameter** sind hier wichtig? What is the **JC69 model**? Why is "JC69" considered to be **the most constrained model of evolution**?

The **JC69 model** is the **simplest substitution model**.

It's **two main assumptions** are:
(1) It assumes **equal base frequencies**
(2) It assumes **equal mutation rates**

The Jukes-Cantor model assumes **equal base frequencies and equal mutation rates**, **therefore it does not have any free parameter**.

There exist two very useful matrices, **BLOSUM** and **PAM**.

The information content in **BLOSUM** and **PAM** matrices is given in terms of "***relative entropy***". The higher the entropy, the higher the evolutionary distance between sequences where scores were derived from.

:: **Statistik**. Erkläre die **statistische Signifikanz** (**p-wert**, Unterschied zwischen **Signifikanz** und **Relevanz**). Aussagen zu Nullhypothese annehmen oder verwerfen. Was sind **Fehler 1. und 2. Art**? 2 Verteilung mit Signifikanzniveau und 1-beta einzeichnen

Der **p-Wert** wird so genannt da er im englischen **p-value** genannt wird, und dies von "**probability**" kommt.

Mit dem **p-Wert** wird angedeutet, **wie extrem ein Ergebnis ist**.

:: Define **genetic distance**. How can it be estimated? What is **observed** vs. **estimated genetic distance**?

:: Was ist **comparative Genomics**? Wie und wofür ist sie anwendbar? Nennen Sie **4 Einsatzmöglichkeiten der Genomanalyse**! [**DONE**]

Comparative genomics is **the study of the relationship of genome structure and function across different biological species**.
Comparative genomics promises to yield insights into many aspects of the evolution of modern species.

Comparative Genomics is based on:
→ **genome mapping** and **sequencing**.
→ **comparing structure of known genes and genome**.

This thus allows us to use a gene **from one species**, and **find a related gene in another species**.

→ Verwendung der **Genome** von **Modellorganismen**:
• **Identifikation** von Genen unbekannter Funktion in **anderen Organismen**
• **Gencluster** und **regulatorische Interaktionen** sind oftmals speziesübergreifend **konserviert**.
Beispiel: Alle wichtigen human- oder pflanzenpathogenen Mikroorganismen werden in den nächsten Jahren sequenziert.
• **Identifikation** der entsprechenden fehlenden bzw. gemeinsamen Gene.
• Vergleich mit **apathogenen Mikroorganismen**.
• **Identifikation** von **Pathogenitäts**- und **Virulenzfaktoren** bzw. von essentiellen „Houskeeping Genen".
• **Identifikation von pharmazeutischen Targets. Vergleichende Genomanalyse ist erst im Entwicklungsstadium!**

Auf **Englisch**:
• Identification of the respective missing or common genes.
• Comparison with **apathogenic micro-organisms**.
• Identification of **pathogenicity** and virulence **factors** or of essential „housekeeping genes ".
• Then, **Identification of pharmaceutical targets**.

Comparative Genomics versucht die Funktion von Genen in „**einfachen**“ Organismen zu erforschen und die Erkenntnisse dann auf andere Organismen (z.B. Mensch) **umzulegen**. Die DNA-Sequenz eines bekannten Gens wird dann **zur Suche nach ähnlichen Teilsequenzen** (Ähnlichkeitssuche) in der neu sequenzierten DNA verwendet.
Mensch: Keine „Gen knock-outs“, keine kontrollierte Züchtung, keine Mutantenkollektion
Modellorganismen: Maus, Ratte, Hamster, Zebrafisch, Hefe.
Programme: **COG** (Cluster of Orthologous Groups), **PEDANT** (Protein Extraction Description and Analysis Tool), ERGO (formelry known as **WIT** database, What Is There).

:: Which **problem** may exist when we **annotate** a genome? [**DONE**]

→ **Errors** in a database can be continued and potentiated.
→ Automatic annotation methods are better suitable for prokaryotes because they have a much simpler genome structure; simple algorithms give better results.

:: Welche **Hindernisse** gibt es bei der **Genomsequenzierung**?

(1) **Lücken** in der Überdeckung
(2) **Fehler** beim Kopieren (Klonen) und beim Sequenzieren: Fehlerquote 1-2 %
(3) **Repeats unterschiedlicher Länge**: Repeats sind Teile der DNA, die wiederholt auftreten. Man spricht auch von Repeats, wenn sich die Teilsequenzen sehr ähnlich sind.

:: Was gilt beim "**Occams Razor**"? [**DONE**]

Es sollte bei **widersprüchlichen Modellen** stets das **einfachere** gewählt werden.

Im **englischen Original**: *“The Occams Razor states that among competing hypotheses, the hypothesis with the fewest assumptions should be selected.”*

Mit anderen Worten sollte man immer zu **einfacheren Theorien** gelangen. Dabei muss die einfachste Theorie nicht unbedingt die genauste sein.

:: Nennen Sie **5 Modellorganismen**. [DONE]
(1) Mouse (*mus musculus*)
(2) **Drosophila melanogaster** (die *schwarzbäuchige Taufliege*)
(3) **Arabidopsis thaliana**
(4) Yeast (*Saccharomyces cerevisiae*)
(5) Schimmelpilz ***Neurospora crassa***
(6) **Danio rerio** (*Zebrabärbling*)

:: Was ist ein “**Dome Plot**”?

:: Nennen Sie **4 Einsatzmöglichkeiten** der **Genomanalyse!**

(1) **Pharmakologische Tests** auf **Genbasis**
(2) **Abstammungsnachweise**
(3) **Verwandtschaftsnachweise**
(4) **Untersuchung** von **metabolischen Netzwerken**

:: Was ist “**Homologie**”, was ist “**Homoplasie**”? Wie kann man man beide unterscheiden? Wieso suchen wir überhaupt ähnliche oder homologe Sequenzen?

In der **Regel** implizieren hohe Sequenzähnlichkeit ähnliche Funktion oder Struktur.

Als **Homologie** bezeichnet man in der biologischen Systematik die grundsätzliche Übereinstimmung von Körperstrukturen zweier Taxa aufgrund ihres **gemeinsamen evolutionären Ursprungs**.

Es gibt **3 Homologie-Kriterien**:
→ Kriterium der **Lage**
→ Kriterium der **spezifischen Qualität**
→ Kriterium der **Kontinuität** (mittels Übergangsreihen)

Im Gegensatz zur Homologie steht der Ausdruck Homoplasie.

Eine **Homoplasie** (griech. Homos = „gleich“ , plasis = „Formung“) bezeichnet in der Biologie ein Merkmal, das bei mehreren unterschiedlichen Taxa **unabhängig** voneinander entstanden ist (=**Konvergenz**). **Homoplasie** umfasst neben analoger Entwicklung außerdem eine **parallele Evolution** (z. B. parallele Entwicklung eines Streifenmusters mit unterschiedlicher molekularer Grundlage aus einem ungestreiften Vorfahren bei zwei Fischarten).

Unterscheiden kann man das ganze anhand der Sequenzinformation.

→ http://de.wikipedia.org/wiki/Homologie_%28Biologie%29

:: Was ist ein “**clade**”, was ist “**cladistic**”? Definieren Sie diese Begriffe.

A **clade** (defined in 1957) is a group consisting of an **ancestor** and all its **descendants**, a single "branch" on the "tree of life".

The ancestor may be an individual, a population or even a species.
Many familiar groups, rodents and insects for example, are **clades**; others, like lizards and monkeys, are not. → http://en.wikipedia.org/wiki/Clade

Anderes Wort für **Cladogramm** ist "**family tree**" - ein **Stammbaum**.

:: **Berechnen** Sie die **Levenshteindistanz** zwischen **TATAATAGA** und **TATATAGA**, wobei Sie als Kosten **1 für eine Ersetzung** und **2 für Einsetzung** oder **Löschen** verwenden. **Initialisieren** Sie die Matrix mit 0,2,4,6,... (Diese Frage ist allgemein in der Form gehalten, das wir die Levenstheindistanz zwischen 2 Sequenzen ausrechnen müssen sowie eine NxM Matrix zeichnen müssen, inklusive "**Backtracking**". Berechnen Sie die **Levensthein-Distanz** zwischen folgenden Zeichenketten: A: CATAATAG, B: CATATAG.

Definieren wir zuerst die **Levenshtein-distance**.
→ This **distance** is a **string metric** for measuring the difference between two sequences.

Informally, the **Levenshtein distance** between two words is **the minimum number** of **single-character edits** (insertion, deletion, substitution) required to change one word into the other.

The phrase **"edit distance"** is often used to refer specifically to **Levenshtein distance**.
It is named after ***Vladimir Levenshtein***, who considered this distance in **1965**.
It is **closely** related to **pairwise string alignments**.

A specific example:

The **Levenshtein-distance** between "**kitten**" and "**sitting**" ist:

→ **kitten** (6 characters)
→ **sitting** (7 character)

We have to insert **one** character ("***insert***") which is the letter **g**. Two more characters must be changed/modified ("***substitution***"), exactly between the letters **k** and **e**. Thus, we require at least **3** "***edit changes***".