Lecture notes on Bioinformatics 2023

Kasper Munch

Contents

1 Preface

What is bioinformatics

Bioinformatics is an interdisciplinary field that combines biology, computer science, mathematics, and statistics to analyze and interpret biological data, particularly large-scale molecular and genetic information. It involves the application of computational methods and techniques to understand biological processes, enhance our knowledge of genetics, and provide insights into various biological phenomena. With advancements in biotechnology, particularly in genomics, transcriptomics, proteomics, and metabolomics, enormous amounts of biological data are generated. These datasets contain information about genes, proteins, molecules, and their interactions. Typical work for a bioinformatician involves data mining and analysis and applying statistical and machine learning techniques to identify patterns, correlations, and significant features in biological datasets.

Bioinformatics involves the development of databases and software tools to efficiently store, organize, and manage these vast datasets. These resources enable researchers to access and manipulate the data for analysis. One of the core aspects of bioinformatics is the analysis of DNA, RNA, and protein sequences. This includes tasks such as sequence alignment, where sequences are compared to identify similarities and differences. Sequence alignment is crucial for understanding evolutionary relationships, identifying functional elements, and detecting genetic variations.

Structural bioinformatics predicts and analyzes the three-dimensional structures of proteins, RNA, and other molecules. Understanding the structure of biomolecules is essential for comprehending their functions, interactions, and mechanisms. To this end, bioinformaticians also build tools used to predict the functions of genes, proteins, and other biomolecules. This involves comparing sequences to known functional elements or domains and inferring their roles based on similarities.

Bioinformaticians also develop and use algorithms to cluster sequences and construct trees revealing the history and relationships among different species or genes. A related area is the comparison of genomes across different species, revealing insights into genomic evolution, gene conservation, and functional divergence.

In medical research, bioinformatics contributes to drug development by predicting potential drug targets, simulating molecular interactions, and identifying candidate compounds for further experimental testing. Bioinformaticians contribute analyses

of individuals' genetic and molecular data and thus support personalized medicine, where medical treatments and interventions can be tailored to a person's unique genetic makeup.

In essence, bioinformatics provides the tools and methodologies to extract meaningful insights from biological data, advancing our understanding of life sciences, genetics, and various other areas of biology. It is a rapidly evolving field that continues to play a pivotal role in modern biological research and applications.

Why should I care?

You may wonder how you find yourself with lecture notes on bioinformatics. The reason is that programming and digital literacy are increasingly important for molecular biologists because of the growing reliance on data-driven approaches, high-throughput technologies, and computational analyses in modern biological research. Programming skills and digital literacy empower molecular biologists to efficiently handle data and perform complex analyses. As biological research becomes more intertwined with computational approaches, these skills have become integral for conducting impactful and innovative research in molecular biology. Here are several reasons why these skills are essential for molecular biologists:

Data Handling and Analysis: Molecular biology research generates vast amounts of data from techniques like DNA sequencing, gene expression profiling, and protein structure determination. Programming skills enable scientists to process, analyze, and extract meaningful insights from these large datasets using computational tools and algorithms.

Efficiency and Automation: Repetitive tasks like data preprocessing can be automated using programming scripts. This increases efficiency and reduces the chances of human error, allowing researchers to focus more on the scientific interpretation of results.

Customized Data Analysis: Pre-existing software tools may not always meet specific research needs. With programming skills, molecular biologists can develop custom scripts and algorithms tailored to their experiments, ensuring optimal analysis and interpretation of results.

Statistical Analysis: Many biological experiments require statistical analysis to draw meaningful conclusions. Programming allows researchers to implement statistical methods, conduct hypothesis testing, and visualize results effectively.

Collaboration and Communication: Digital literacy allows researchers to effectively collaborate by sharing data, code, and results online. This promotes transparency, reproducibility, and knowledge sharing within the scientific community.

Keeping Up with Advancements: Many new technologies and techniques in molecular biology are heavily reliant on computational analyses. Researchers with programming skills can easily adapt to new methodologies and stay current with the rapidly evolving field.

Reading this book

In this course, we do not cover all aspects of bioinformatics but focus on aspects that relate to the analysis of genes and genomes. These topics will become increasingly dominant as personalized medicine becomes the norm and introduces many essential bioinformatics concepts, models, and algorithms. Many of you follow the course Biological Structure and Function, teaching structural biology in bioinformatics.

In each of the topics we cover, we will spend the lectures building an understanding of the algorithms and inner workings of the relevant bioinformatics tools. We will use the exercises to test out some of the tools in small self-contained bioinformatics projects. # Databases and resources

Bioinformatics databases are crucial resources that provide access to a vast array of biological data, ranging from genetic sequences and protein structures to functional annotations and disease-related information. These databases play a pivotal role in various research fields, aiding scientists in data analysis, hypothesis testing, and discovery. Here's an overview of some of the most important bioinformatics databases and how to use them:

GenBank: GenBank is a comprehensive database maintained by the National Center for Biotechnology Information (NCBI) that contains DNA sequences submitted by researchers worldwide. It includes sequences from various organisms, along with associated metadata and annotations. Users can search for specific sequences, access genome assemblies, and retrieve information for various genomic regions.

Key Features of GenBank:

Sequence Collection: GenBank contains a vast collection of DNA, RNA, and protein sequences from diverse organisms, including viruses, bacteria, fungi, plants, animals, and more.

Annotations: Sequences in GenBank are accompanied by annotations that provide valuable information about the sequence, such as gene names, protein products, functional regions, and other relevant data.

Genomic Assemblies: GenBank hosts whole genome assemblies for many organisms, enabling researchers to access and analyze complete genomes.

Versioning: Each sequence in GenBank is assigned a version number to track changes and updates. This allows researchers to access historical versions of sequences.

Cross-References: GenBank cross-references other databases and resources, facilitating integration and collaboration among different data sources.

BioProject and BioSample: GenBank integrates with BioProject and BioSample databases, providing additional context for submitted sequences, such as information about the source organism and experimental details.

Search and Retrieval: The GenBank website offers a user-friendly interface for searching and retrieving sequences based on keywords, accession numbers, organisms, and more.

UniProt: UniProt is a database that provides detailed information about protein sequences and their functional annotations. It integrates data from various resources, offering a centralized repository for protein-related information. Researchers can search for proteins by name, sequence, or keywords and access data such as protein function, structure, domains, and post-translational modifications.

UniProt (Universal Protein Resource) is a comprehensive and widely used bioinformatics database that provides a centralized repository of protein sequences and functional information. Maintained collaboratively by the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR), UniProt serves as a valuable resource for researchers, educators, and students in the fields of biology, genetics, and molecular biology.

Key Features of UniProt:

Protein Sequences: UniProt contains a vast collection of protein sequences from a wide range of organisms, including bacteria, archaea, fungi, plants, animals, and viruses.

Functional Annotations: Each protein entry in UniProt is accompanied by extensive functional annotations, providing information about the protein's biological role, structure, domains, post-translational modifications, interactions, subcellular localization, and more.

Cross-References: UniProt integrates information from other databases and resources, offering cross-references to relevant data sources, such as gene databases, literature citations, and 3D protein structures.

Isoforms and Variants: UniProt includes information about protein isoforms, splice variants, and natural variants that affect protein function.

Taxonomic Information: Protein entries are categorized based on taxonomy, enabling users to explore protein diversity across different organisms.

Sequence Alignments: UniProt provides sequence alignments for families of homologous proteins, aiding in understanding protein evolution and relationships.

Disease Annotations: UniProt includes annotations linking proteins to various diseases, such as genetic disorders and cancers, based on experimental evidence.

Literature Citations: Protein entries often include references to scientific literature, allowing users to access the original research articles related to specific proteins.

Keywords and Annotations: UniProt uses controlled vocabularies to assign keywords and annotations to proteins, facilitating standardized and structured data representation.

Components of UniProt:

UniProt is divided into three main components:

UniProtKB: This is the primary component, providing a comprehensive collection of protein sequences and annotations. UniProtKB is further subdivided into three subdatabases:

UniProtKB/Swiss-Prot: Curated protein entries with high-quality annotations. UniProtKB/TrEMBL: Automatically annotated protein entries that have not yet been curated. UniRef: This component clusters closely related protein sequences into UniRef clusters, reducing redundancy and simplifying sequence analysis.

UniParc: UniParc contains a unique and non-redundant collection of all publicly available protein sequences, regardless of their level of annotation.

Usage of UniProt:

To use UniProt effectively:

Search: Enter protein names, identifiers, gene names, keywords, or organisms in the search bar. Browse: Explore protein entries by taxonomy, organism, keywords, and functional annotations. Retrieve Information: Access protein sequences, functional annotations, cross-references, and literature citations. Analyze: Utilize the provided data to understand protein function, structure, interactions, and disease associations. Integration: Integrate UniProt data with other bioinformatics tools and resources for comprehensive analyses. UniProt is an invaluable resource for researchers studying proteins, their functions, and their roles in various biological processes. Its user-friendly interface, extensive annotations, and cross-referencing capabilities make it a cornerstone tool in the field of molecular biology.

PubMed: PubMed is a database maintained by the National Library of Medicine, offering access to an extensive collection of biomedical literature. Researchers can search for articles related to specific topics, diseases, genes, or proteins. PubMed provides abstracts, full-text articles, and links to external resources, aiding literature review and research planning.

Ensembl: Ensembl is a genome browser and annotation database that provides genomic sequences, gene annotations, and functional information for a wide range of species. Users can explore gene structures, regulatory elements, genetic variations, and comparative genomics data. Ensembl's browser interface allows for interactive exploration of genomic regions.

NCBI Gene: This database offers comprehensive information about genes, including functional annotations, expression data, gene-disease associations, and orthologous relationships. Researchers can search for specific genes, access genetic and protein sequences, and retrieve details about gene function and expression patterns.

The NCBI Gene database, maintained by the National Center for Biotechnology Information (NCBI), is a comprehensive resource that provides information about genes from a wide range of organisms. It serves as a centralized repository for gene-related data, annotations, and functional information. The database plays a crucial role in molecular biology research, aiding scientists in understanding gene functions, interactions, and regulatory mechanisms.

Key Features of the NCBI Gene Database:

Gene Information: The NCBI Gene database contains detailed information about individual genes, including gene names, symbols, aliases, genomic locations, chromosomal coordinates, and orientations.

Gene Function: Each gene entry is accompanied by functional annotations that describe the gene's biological roles, molecular functions, and involvement in cellular processes.

Transcripts and Isoforms: The database provides information about different transcript variants and isoforms associated with a gene. This includes details about alternative splicing, coding regions, and untranslated regions.

Orthologs and Paralogs: NCBI Gene includes information about orthologous and paralogous genes across different species, aiding in understanding gene evolution and relationships.

Homology and Alignment: Gene entries often include sequence alignments, similarity scores, and evolutionary relationships with related genes.

Protein Products: NCBI Gene provides information about the protein products encoded by genes, including functional domains, post-translational modifications, and protein interactions.

Gene Ontology (GO) Annotations: Many gene entries are associated with Gene Ontology terms that categorize gene functions, cellular components, and biological processes.

Expression Data: The database offers data on gene expression patterns in various tissues, developmental stages, and conditions, aiding in understanding gene regulation.

Genomic Context: Gene entries include information about neighboring genes, regulatory elements, and genetic variations in the genomic vicinity.

Disease Associations: Gene entries may include annotations linking genes to specific diseases, based on experimental evidence.

Usage of the NCBI Gene Database:

To use the NCBI Gene database effectively:

Search: Enter gene names, symbols, IDs, or keywords in the search bar. Browse: Explore gene entries by species, functional annotations, and genomic locations. Retrieve Information: Access gene details, functional annotations, orthologs, paralogs, and protein products. Analyze: Utilize the provided data to understand gene functions, expression patterns, and disease associations. Cross-Reference: NCBI Gene integrates with other NCBI databases, providing links to nucleotide sequences, protein sequences, PubMed articles, and more. The NCBI Gene database serves as a critical resource for researchers studying individual genes and their roles in various biological processes. Its comprehensive annotations, integration with other NCBI resources, and user-friendly interface make it an essential tool for molecular biologists, geneticists, and researchers across diverse fields.

dbSNP: The Single Nucleotide Polymorphism Database (dbSNP) is a repository of genetic variations, including SNPs and other types of variations. Researchers can search for specific variations, retrieve their annotations, and explore data related to population frequencies, diseases, and functional effects.

The dbSNP (Single Nucleotide Polymorphism Database) is a comprehensive bioinformatics resource maintained by the National Center for Biotechnology Information (NCBI). It serves as a repository for genetic variations, including single nucleotide polymorphisms (SNPs), insertions, deletions, and other types of genetic variants. dbSNP plays a crucial role in cataloging and providing access to genetic variations across various species, aiding researchers in understanding the genetic basis of traits, diseases, and population diversity.

Key Features of the dbSNP Database:

Variation Catalog: dbSNP contains a wide variety of genetic variations, including SNPs, short insertions and deletions (indels), microsatellites, and larger structural variants.

Variant Annotations: Each genetic variant is accompanied by annotations, such as genomic coordinates, alleles, frequencies in different populations, and clinical significance (if applicable).

Population Data: dbSNP provides information about the frequency of genetic variants in different human populations, enabling researchers to study population genetics and diversity.

Functional Annotations: Many genetic variants are annotated with potential functional consequences, such as effects on protein coding, splicing, or regulatory elements.

Cross-References: dbSNP cross-references other databases, allowing users to access additional information about genes, proteins, and diseases associated with specific genetic variants.

Genomic Context: The database includes information about the genomic context of variations, such as neighboring genes, regulatory elements, and conserved regions.

ClinVar Integration: dbSNP integrates with the ClinVar database, providing information about the clinical significance of genetic variants and their associations with diseases.

Genome Build Compatibility: dbSNP supports different genome builds, enabling users to map variations to specific versions of the genome.

Usage of the dbSNP Database:

To use the dbSNP database effectively:

Search: Enter SNP IDs, genomic coordinates, gene names, or keywords in the search bar. Browse: Explore variants by chromosome, population, frequency, or functional annotations. Retrieve Information: Access variant details, genomic coordinates, allele frequencies, and clinical significance information. Analyze: Utilize variant annotations for studying disease associations, population genetics, and functional effects. Integration: Combine dbSNP data with other resources to enhance genetic and genomic analyses. dbSNP is a crucial resource for researchers investigating the genetic variations that contribute to phenotypic differences, diseases, and population diversity. Its extensive collection of annotated variants, integration with other databases, and user-friendly interface make it an indispensable tool for geneticists, epidemiologists, and researchers in related fields.

OMIM: Online Mendelian Inheritance in Man (OMIM) is a database that catalogs genetic mutations and their relationships to inherited diseases. Researchers can search for genes, phenotypes, and diseases to access information about genetic disorders, associated mutations, and relevant literature.

The Online Mendelian Inheritance in Man (OMIM) database is a comprehensive and authoritative resource that catalogues information about genetic disorders, traits, and other phenotypic traits in humans. Maintained by the McKusick-Nathans Institute of Genetic Medicine at Johns Hopkins University, OMIM plays a pivotal role in bridging the gap between genetic research and clinical medicine, providing valuable insights into the genetic basis of various human conditions.

Key Features of the OMIM Database:

Genetic Disorders: OMIM contains detailed information about a wide range of genetic disorders, both rare and common, providing insights into the genetic mutations, inheritance patterns, clinical features, and molecular mechanisms underlying these conditions.

Phenotypic Traits: In addition to genetic disorders, OMIM also catalogues information about various phenotypic traits, such as physical characteristics, susceptibility to diseases, and responses to treatments.

Gene-Centric Information: OMIM provides gene-centric entries that include information about specific genes associated with genetic disorders. These entries detail gene functions, mutations, protein products, and any known associations with diseases or traits.

Clinical Descriptions: Each entry includes detailed clinical descriptions of the disorders or traits, including symptoms, diagnostic criteria, and information about disease progression.

Genetic Inheritance Patterns: OMIM categorizes disorders based on their inheritance patterns, such as autosomal dominant, autosomal recessive, X-linked, and mitochondrial.

Gene and Locus References: OMIM cross-references genes and loci to other databases and resources, facilitating integration with genomic and functional data.

Links to Literature: OMIM entries include references to scientific literature, providing access to primary research articles and reviews related to specific disorders and genes.

Molecular Mechanisms: Many OMIM entries describe the molecular mechanisms underlying genetic disorders, providing insights into how mutations affect biological processes.

Usage of the OMIM Database:

To use the OMIM database effectively:

Search: Enter disease names, gene names, phenotypic keywords, or gene identifiers in the search bar. Browse: Explore entries by disease categories, gene names, inheritance patterns, and other criteria. Retrieve Information: Access detailed information about genetic disorders, phenotypic traits, genes, clinical descriptions, and molecular mechanisms. Analyze: Utilize the provided data to understand the genetic basis of diseases, inheritance patterns, and potential therapeutic targets. Integration: Combine OMIM data with other genetic and genomic resources to enhance research and clinical applications. OMIM is widely used by geneticists, clinicians, researchers, and healthcare professionals to gain insights into the genetic underpinnings of diseases, guide clinical diagnoses, and inform therapeutic strategies. Its rich content, authoritative sources, and user-friendly interface make it an essential tool for anyone involved in genetics and genomics research.

STRING: The STRING database provides information about protein-protein interactions and functional associations. Users can search for a protein of interest to explore its interactions, view network maps, and analyze the potential functions of proteins within a network context.

To use these databases effectively:

Keyword Search: Start by entering relevant keywords, gene names, protein identifiers, or disease names in the search bar. Filters: Many databases offer filters to refine search

results based on criteria such as species, data type, and publication date. Advanced Search: Use advanced search options to specify specific fields or search criteria to narrow down results. Data Retrieval: Once you find relevant information, you can often download sequences, annotations, and other data for further analysis. Citations: When using data from these databases in your research, make sure to provide proper citations to acknowledge the original sources. These databases are continuously updated and expanded, so it's essential to explore their documentation and tutorials to stay informed about the latest features and search strategies.

The STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) is a valuable bioinformatics resource that provides information about protein-protein interactions (PPIs) and functional associations. STRING helps researchers explore the relationships between proteins, gain insights into biological processes, and understand the complex networks that underlie various cellular functions. Maintained by a collaborative team of researchers, STRING is widely used in various fields, including molecular biology, systems biology, and drug discovery.

Key Features of the STRING Database:

Protein Interaction Data: STRING aggregates and integrates experimental and computational data on protein interactions from diverse sources, including high-throughput experiments, curated databases, and literature mining.

Functional Associations: Beyond direct physical interactions, STRING provides information about functional associations between proteins, such as co-expression, shared domains, and similar annotations.

Confidence Scores: STRING assigns confidence scores to interactions and associations based on the evidence supporting them. These scores help users assess the reliability of the reported interactions.

Network Visualization: STRING generates interactive graphical representations of protein networks, allowing users to visualize and explore protein interactions and functional relationships in a visual format.

Species Coverage: STRING supports a wide range of organisms, from model organisms to less-studied species, enabling researchers to investigate protein interactions in different biological contexts.

Enrichment Analysis: STRING offers enrichment analysis tools that help users identify overrepresented functional categories, pathways, and Gene Ontology terms within a set of proteins of interest.

Prediction Methods: STRING includes computational prediction methods that estimate potential interactions based on sequence similarities, domain architectures, and other features.

Cross-References: STRING provides cross-references to external resources, allowing users to access additional information about interacting proteins, pathways, and diseases.

Usage of the STRING Database:

To use the STRING database effectively:

Search: Enter protein names, identifiers, or keywords to find interactions and associations. Network Visualization: Explore and visualize protein interaction networks, adjusting confidence thresholds and interaction types. Retrieve Information: Access interaction scores, functional annotations, pathway information, and external references. Analyze: Utilize the provided data to understand protein functions, regulatory mechanisms, and disease associations. Enrichment Analysis: Perform enrichment analysis to identify functional themes within a set of proteins. Integration: Combine STRING data with other omics data to gain a holistic understanding of cellular processes. STRING is a versatile tool for researchers investigating protein interactions, cellular pathways, and systems-level biology. Its user-friendly interface, comprehensive data sources, and advanced analysis features make it an essential resource for studying the complex molecular networks that govern biological functions.

Sequence databases

Genome browsers

The UCSC Genome Browser is a widely used and comprehensive online tool for visualizing and exploring genomic information from a wide range of organisms. Developed and maintained by the University of California, Santa Cruz (UCSC), this browser offers a user-friendly interface to navigate and analyze genomes, gene structures, regulatory elements, genetic variations, and other genomic features. Here's an overview of the UCSC Genome Browser and its key features:

Features of the UCSC Genome Browser:

Genome Selection: The UCSC Genome Browser supports a vast collection of genomes from various species, including humans, model organisms, and microbes. Users can select the genome of interest from a dropdown menu.

Genome Views: The browser provides different views of the genome, including the standard linear view, a circular view (for some genomes), and a conservation view that highlights evolutionarily conserved regions.

Navigation and Zoom: Users can easily navigate through genomic regions using navigation arrows and zoom controls. This allows detailed examination of specific regions.

Annotation Tracks: One of the most powerful features of the UCSC Genome Browser is the ability to overlay various types of annotation tracks onto the genome view. These tracks include gene annotations, regulatory elements, genetic variations, and more. Users can customize which tracks are displayed and adjust their visibility and order.

Gene Annotations: The browser displays gene structures, including exons, introns, coding sequences, and untranslated regions. It provides information about gene names, functions, and alternative splicing patterns.

Regulatory Elements: Regulatory elements such as promoters, enhancers, and transcription factor binding sites can be visualized, aiding in understanding gene regulation.

Genetic Variations: Genetic variations, including single nucleotide polymorphisms (SNPs), insertions, and deletions, can be displayed. Users can examine variations' positions, alleles, frequencies, and potential functional consequences.

Custom Tracks: Users can upload their own data, such as experimental results or custom annotations, to visualize alongside the provided tracks.

Comparative Genomics: The browser offers tools to compare genomic sequences and annotations across multiple species. This aids in identifying conserved regions and evolutionary changes.

Search and Highlight: Users can search for specific genes, regions, sequences, or annotations and highlight them in the genome view for detailed examination.

Links to External Resources: The UCSC Genome Browser provides links to external databases and resources, enabling seamless access to additional information.

Usage:

To use the UCSC Genome Browser:

Access the Browser: Go to the UCSC Genome Browser website (genome.ucsc.edu).

Choose a Genome: Select the genome of interest from the dropdown menu.

Navigate and Explore: Navigate to specific genomic regions using the search bar, navigation arrows, and zoom controls. Customize the displayed tracks to focus on relevant features.

Analyze and Download: Analyze gene structures, regulatory elements, and genetic variations in the context of the genome. Download data and images for use in research or presentations.

The UCSC Genome Browser serves as an invaluable resource for researchers, educators, and students seeking to understand and analyze genomic information across various organisms. Its intuitive interface, extensive collection of annotation tracks, and powerful visualization capabilities make it a cornerstone tool in the field of genomics.

list of databases

GenBank UniProt PubMed Ensembl NCBI Gene STRING dbSNP OMIM PDB (Protein Data Bank) KEGG (Kyoto Encyclopedia of Genes and Genomes) Reactome GO (Gene Ontology) HPRD (Human Protein Reference Database) COSMIC (Catalogue Of Somatic Mutations In Cancer) ClinVar TCGA (The Cancer Genome Atlas) Pfam Rfam InterPro GTEx (Genotype-Tissue Expression Project) HGNC (HUGO Gene Nomenclature Committee) FlyBase WormBase TAIR (The Arabidopsis Information Resource) Mouse Genome Informatics (MGI) RGD (Rat Genome Database) PharmGKB DrugBank PubChem dbGaP (Database of Genotypes and Phenotypes) dbCAN (Carbohydrate-Active enZYmes Database) CATH (Class, Architecture, Topology, Homology) MEROPS (Peptidase Database) BioGRID IntAct IUPHAR/BPS Guide to Pharmacology GenAtlas miRBase lncRNAdb DGV (Database of Genomic Variants) UCSC Genome Browser ExPASy (Expert Protein Analysis System) Swiss-Model Pfam SUPERFAMILY EMBL-EBI InterProScan WikiPathways ChEMBL BioCyc

1. GenBank
2. UniProt
3. PubMed
4. Ensembl
5. NCBI Gene
6. STRING
7. dbSNP
8. OMIM
9. PDB (Protein Data Bank)
10. KEGG (Kyoto Encyclopedia of Genes and Genomes)
11. Reactome
12. GO (Gene Ontology)
13. HPRD (Human Protein Reference Database)
14. COSMIC (Catalogue Of Somatic Mutations In Cancer)
15. ClinVar
16. TCGA (The Cancer Genome Atlas)
17. Pfam
18. Rfam
19. InterPro
20. GTEx (Genotype-Tissue Expression Project)
21. HGNC (HUGO Gene Nomenclature Committee)
22. FlyBase
23. WormBase
24. TAIR (The Arabidopsis Information Resource)
25. Mouse Genome Informatics (MGI)
26. RGD (Rat Genome Database)
27. PharmGKB

28. DrugBank
29. PubChem
30. dbGaP (Database of Genotypes and Phenotypes)
31. dbCAN (Carbohydrate-Active enZYmes Database)
32. CATH (Class, Architecture, Topology, Homology)
33. MEROPS (Peptidase Database)
34. BioGRID
35. IntAct
36. IUPHAR/BPS Guide to Pharmacology
37. GenAtlas
38. miRBase
39. lncRNAdb
40. DGV (Database of Genomic Variants)
41. UCSC Genome Browser
42. ExPASy (Expert Protein Analysis System)
43. Swiss-Model
44. Pfam
45. SUPERFAMILY
46. EMBL-EBI InterProScan
47. WikiPathways
48. ChEMBL
49. BioCyc
50. dbVar

Knowledege data bases

Do it yourself:

2 Genome-wide association studies

Genetic disease association

Common disease common variant hypothesis

The common disease common variant (CDCV) hypothesis is a concept in human genetics that seeks to explain the genetic basis of common complex diseases by focusing on the prevalence of common genetic variants in the population. This hypothesis contrasts with the "rare variant" hypothesis, which suggests that rare genetic variants with larger effects play a significant role in disease susceptibility.

Hypothesis Explanation:

The CDCV hypothesis suggests that the genetic susceptibility to many common complex diseases, such as diabetes, heart disease, and certain types of cancer, is primarily influenced by common genetic variants that are present at relatively high frequencies in the population. Common genetic variants are those that occur in more than 1% of the population. These variants are typically single nucleotide polymorphisms (SNPs), where a single DNA base is substituted with another.

According to this hypothesis, multiple common genetic variants, each with a small effect size, collectively contribute to an individual's risk of developing a specific disease. The idea is that these common variants, which are widely distributed in the population, are responsible for a significant portion of the genetic susceptibility to a given disease.

Key Points of the CDCV Hypothesis:

Polygenic Nature: Common complex diseases are often influenced by multiple genetic factors, each with a modest effect size. This polygenic nature suggests that no single genetic variant has a large impact on disease risk.

Complex Inheritance: Common variants contribute to disease risk in a complex manner, influenced by interactions between genes and the environment. This complexity makes it challenging to predict disease risk solely based on genetic information.

Quantitative Trait Loci (QTLs): Common variants associated with disease risk are also often associated with variations in quantitative traits, such as blood pressure or cholesterol levels, which are risk factors for certain diseases.

Genome-Wide Association Studies (GWAS): The CDCV hypothesis has been supported by the findings of genome-wide association studies (GWAS). These studies scan the entire genome for associations between common genetic variants and specific diseases or traits. Many common variants have been identified as having modest effects on disease risk through GWAS.

Statistical Significance: Given the small effect sizes of individual variants, large sample sizes are often required in GWAS to achieve statistical significance. This emphasizes the cumulative impact of multiple variants on disease risk.

Heritability Estimates: The CDCV hypothesis aligns with estimates of heritability for common complex diseases, which suggest that a significant portion of the variation in disease risk can be attributed to genetic factors.

Critiques and Challenges:

While the CDCV hypothesis has provided valuable insights into the genetic basis of common complex diseases, it is not without its criticisms and challenges. Some researchers argue that rare genetic variants with larger effect sizes could also contribute to disease risk and might have been overlooked by traditional GWAS approaches.

In recent years, advancements in sequencing technologies and analysis methods have allowed researchers to explore the role of rare variants in disease susceptibility, blurring the distinction between the CDCV hypothesis and the "rare variant" hypothesis. It's increasingly recognized that both common and rare variants contribute to the overall genetic architecture of complex diseases, and a comprehensive understanding of disease risk requires considering both types of variants.

Testing association

Genome-wide association

Introduction:

The completion of the Human Genome Project in 2003 marked a pivotal milestone in the realm of genomics, providing a comprehensive map of human genetic information. Subsequent advancements in high-throughput genotyping and sequencing technologies have ushered in an era of genome-wide association studies (GWAS), enabling researchers to unravel the complex genetic underpinnings of various traits, diseases, and phenotypic variations. Genome-wide association studies have revolutionized our understanding of the genetic basis of multifactorial traits, providing insights into the interplay between genetics, environment, and disease susceptibility. This introduction

aims to provide an academic overview of GWAS, elucidating their methodology, significance, challenges, and potential implications in the fields of genetics, medicine, and personalized healthcare.

Methodological Underpinnings of GWAS: Genome-wide association studies are designed to identify statistically significant associations between genetic variants and phenotypic traits within a population. These genetic variants, typically single nucleotide polymorphisms (SNPs), are distributed across the human genome. The fundamental approach of GWAS involves genotyping a large number of individuals for a multitude of SNPs, followed by a statistical analysis to identify genetic variants that are overrepresented in individuals with a particular trait or condition. This analysis often involves comparing allele frequencies between cases and controls, taking into account potential confounding factors.

Significance and Implications: The insights garnered from GWAS have provided a remarkable understanding of the genetic basis of various complex traits and diseases, including diabetes, cardiovascular diseases, psychiatric disorders, and cancer susceptibility. By pinpointing specific genetic variants associated with disease risk, GWAS have not only expanded our fundamental knowledge of human genetics but have also identified potential therapeutic targets and pathways that underlie these conditions. Moreover, GWAS findings have paved the way for the development of polygenic risk scores, which offer personalized assessments of disease susceptibility based on an individual's genetic makeup.

Challenges and Limitations: Despite their transformative impact, GWAS are not without challenges. One of the major limitations is the "missing heritability" phenomenon, where the cumulative effect of identified genetic variants often does not account for the entire heritable component of a trait. Additionally, the identified genetic variants might have modest effect sizes, necessitating the analysis of large sample sizes to achieve statistical significance. The issue of population stratification, where differences in genetic ancestry can lead to spurious associations, requires meticulous consideration in study design and analysis.

Future Directions and Implications: As GWAS methodologies continue to evolve, integrating data from diverse populations and incorporating functional genomics approaches hold the promise of unraveling the biological mechanisms underlying identified associations. The emergence of large-scale biobanks, coupled with advancements in computational biology and artificial intelligence, is expected to enhance our ability to predict disease risk, identify novel drug targets, and tailor interventions for individuals based on their genetic profiles.

Conclusion:

Genome-wide association studies have emerged as a cornerstone of contemporary genetic research, providing a robust framework for unraveling the intricate relationship between genetic variation and complex traits. Their contributions to our understanding

of disease etiology, personalized medicine, and the genetic basis of human diversity are undeniable. However, ongoing efforts to address methodological challenges, expand diverse representation, and decipher the functional relevance of identified variants are crucial for fully capitalizing on the potential of GWAS in shaping the future of genetics and healthcare.

Introduction to Methods Used in Genome-Wide Association Studies

Genome-wide association studies (GWAS) have revolutionized the field of genetics and genomics by enabling researchers to uncover the genetic underpinnings of complex traits and diseases. GWAS involve analyzing a vast number of genetic markers across the entire genome to identify associations between specific genetic variants and traits of interest. This introduction provides an overview of the methods used in GWAS and their significance in unraveling the genetic basis of various traits and conditions.

Genetic Variation and Disease Risk: Human genetic variation plays a crucial role in determining an individual's susceptibility to various diseases and traits, ranging from common disorders like diabetes and heart disease to more intricate traits such as height or cognitive abilities. GWAS aims to identify the specific genetic variants that contribute to these traits and understand how they influence disease risk or trait expression.

Marker Selection and Genotyping: In a GWAS, a comprehensive set of genetic markers, typically single nucleotide polymorphisms (SNPs), are selected from across the genome. SNPs are single-letter variations in the DNA sequence that can influence traits. High-throughput genotyping technologies allow researchers to determine the genetic variants present in thousands of individuals simultaneously.

Case-Control or Cohort Design: GWAS often follow either a case-control or cohort design. In a case-control design, individuals with a specific trait or disease (cases) are compared to individuals without the trait or disease (controls) to identify genetic variants associated with the trait. In a cohort design, a large group of individuals is followed over time to observe the relationship between genetic variants and the development of traits or diseases.

Statistical Analysis: GWAS involves rigorous statistical analysis to identify associations between genetic variants and traits. The basic approach is to compare the frequency of each genetic variant between cases and controls. Variants that show significantly different frequencies are considered potentially associated with the trait. Various statistical tests, such as chi-square tests or logistic regression, are employed to assess the strength and significance of these associations.

Population Stratification and Confounding: One challenge in GWAS is population stratification, where differences in genetic ancestry among study participants can lead to false positive associations. To mitigate this, researchers use methods to correct for population structure and confounding factors that might influence the results.

Multiple Testing Correction: Due to the large number of genetic markers tested, there's a risk of identifying false positives by chance alone. Multiple testing correction methods, such as the Bonferroni correction or false discovery rate control, are employed to account for the increased likelihood of false positives and to determine statistical significance thresholds.

Linkage Disequilibrium and Fine-Mapping: Genetic variants in close proximity on the same chromosome often exhibit linkage disequilibrium (LD), meaning they are inherited together more frequently than expected by chance. This property allows researchers to pinpoint the region of the genome where the causal variant likely resides. Fine-mapping techniques aim to narrow down the location of the true causal variant within an associated region.

In conclusion, GWAS represent a powerful approach to uncovering the genetic factors influencing traits and diseases. By combining advanced genotyping technologies, robust statistical methods, and meticulous study designs, researchers can identify key genetic variants associated with various conditions, shedding light on the underlying biological mechanisms and paving the way for more personalized approaches to healthcare and disease prevention. ## Do it yourself: Exploring an association with Parkinsons disease {-}

The purpose of this exercise is to expose you to the different kinds of information that are stored in databases relevant to bioinformatics. It takes experience and skill to navigate the user interface of these databases. Here you will see a few of the important ones, but there are many more. I have put a list at the end of this exercise with some additional relevant ones. Browse them at your harts content.

The MacGuffin of this exercise is Parkinsons disease (PD). Millions of people live with PD, which is a progressive incurable disease affecting the nervous system resulting in shaking, involuntary movement and trouble talking and walking. PD is a complex and multifactorial disease with a heritable component: close relative of people with PD have higher risk. 15% of PD patients have a parent, sibling or child with PD. However, unlike other genetic diseases like cystic fibrosis, where a single gene is responsible, the risk of developing PD is determined by small contributions of many different genes. These genes interact with environmental conditions, such as tobacco smoke. PD symptoms are induced by death certain brain cells in the sunstiantia migraine part of the brain that normally produces the neurotransmitter dopamine.

In the study published in by Do at al. 2011, the authors searched for genes contributing to the risk of developing PD. They genotyped 3,426 PD patients (cases) and 29,624 healthy controls for 522,782 single nucleotide polymorphisms SNPs. A SNP is a position in the genome where the nucleotide differ between individuals in a population because some carry a recent mutation (G) and others carry the ancestral variant (A). Simply put, they considered each of the 522,782 SNPs to see if the occurrence of one of the alleles was significantly more often found among the individuals with PD than among healthy controls. Knowing only the number of cases and controls and the total numbers

of A and G alleles, the expected number of A and G alleles found among cases and controls can be computed (E.g. the expected number of cases with the A allele is $(3426 * 23615)/33050 \approx 2448$):

|---:|---:|---:|---:| | | **A** | **G** | Totals | | *Cases* | **2448** | **978** | 3426 | | *Controls* | **21167** | **8456** | 29624 | | Totals | 23615 | 9434 | 33050 |

This corresponds to an expected frequency of Allele A of 40% in both cases and controls. In the study the they observed the following numbers in each group;

|---:|---:|---:|---:| | | *Allele 1* | *Allele 2* | Totals | | *Cases* | **2320** | **1106** | 3426 | | *Controls* | **21295** | **8328** | 29624 | | Totals | 23615 | 9434 | 33050 |

Corresponding to a 48% frequency of allele A among cases and a 39% frequency among controls. To test the statistical significance of this deviation from the expectation, we can perform a Chi-square testto compute the p-value for the above case. The test statistic is very simple:

$$\sum \frac{(observed - expected)^2}{expected}$$

Under the null-hypothesis, the test statistic follows a chi-square distribution with 1 degree of freedom. You can even do this in Python using the `chi2_contingency` from the `script.stats` library. Try it out and see what the p-value is in this case:

```
from scipy.stats import chi2_contingency
observed = [[2320, 1106],
            [21295, 8328]]
stat, p_value, deg_free, expected = chi2_contingency(observed, correction=False)
print(p_value)
```

The p-value is probably something like 3e-07, which you may think is really small. Normally we use 0.05 as our significance cutoff - the probability of wrongly rejecting our null hypothesis of no association. Since there is a 0.05 probably in every test, we need to divide our significance cutoff by the number of tests to get the significance cutoff needed to only wrongly reject the null-hypothesis with 0.05 probability after doing all 522,782 tests. This cutoff is $0.05/522782 = 9.56e - 08$. This adjustment of the significance cutoff is called Bonferroni correction overcorrects a bit, but the correction shows that even p-values as low as the one you computed above is only just significant in a GWAS.

That is pretty much what GWAS is. There are some additional nuts and bolts required to take biases into account. Accounting for cases and controls not being sampled in the same way. You can probably imagine why the SNPs determining eye color would show up as associated with PD if all cases were sampled in Spain and all controls in Finland. However, we will gloss over those details here.

The SNPs associated with PD are not themselves involved in PD, but they mark a genomic neighborhood that is inherited along with the SNP. The closer a gene is to an associated SNP, the more likely it is that the gene is inherited along with the SNP. So it this genomic neighborhood that we should search for genes playing a role in the development of PD. Such candidate genes must be identified for new insights into causes of the disease and how it can be treated. The following part the of the exercise, but is meant to take you though some of the motions that a scientist would go through to search relevant genes in the region.

- Each known SNP has a unique accession number that identifies it.
- The SNP identified by Do et al. has accession number *rs11868035*
- To explore the genomic neighbourhood of this SNP we use the UCSC genome browser.
- The genome browser saves your preferences across sessions so if you have tried this service before before you can reset the browser to the default settings to align the views with this exercise. Click the Genome Browser->Reset All User Settings option in the top menu bar. However, be aware that this action will remove all custom tracks and will clear all track filter and configuration settings that may have modified.
- Under "Find Position" select the most recent human genome assembly (hg38) and paste in rs11868035 under "Position/Search Term" and click GO.
- On the resulting listing click the top link under "NHGRI-EBI Catalog of Published Genome-Wide Association Studies".
- Familiarise yourself with the browser view.
- Notice the SNP shown in the middle of the "NHGRI-EBI Catalog of Published Genome-Wide Association Studies" track.
- Notice the scale shown at the top of the view. What part of the genome is shown? The chromosome ideogram at the top also shows with a red marker what part of the chromosome is shown. Use the zoom buttons to zoom out so you can see a see roughly 1,000 kilobases (1,000,000 base pairs) of the genome.
- The view is divided into tracks. Mousing over the right margin of each track will highlight it in green, and right-clicking will show customisation settings for each track. By holding and dragging on the grey bars furthest to the right, you can even reorder tracks as you please.
- Scrolling down below the view, you can see a large number of tracks you can enable to show in the view.

- In the "Phenotype and Literature" section find the dropdown labelled OMIM Genes and chose the view mode "Pack". Similarly, in the "Expression" section find and enable the track called GNF Atlas 2. Finally, "Mapping and Sequencing", find the Publications track and choose the "Dense" view mode. Now click any of the "Refresh" buttons to update the view with the selected tracks.
- If you click the grey bar on the right of each track, you are redirected to a page explaining the track.
-

> In part I of the exercise, the UCSC "Genes" track has been replaced by "GENCODE v29". If you do not see the "GENCODE v29" track or the "RefSeq Genes" and "Human mRNAs" tracks, scroll down to the drop-down menus under "Gene and Gene Predictions" and select "GENCODE v29" and "NCBI RefSeq". The "Human mRNAs" track can be found under "mRNA and EST". Unfortunately, the GAD view and the Publications track are no longer available.

Have them read a GWAS review

Part of the exercise could be to make chi-square tests for the subset of SNPs in the region to see if they get something similar to

```
import matplotlib.pyplot as plt

pvalues = [1,2,3]
coordinates = [1,2,1]
plt.scatter(pvalues, coordinates)
plt.title('Plot 1')
plt.show()

pvalues = [1,2,3]
coordinates = [1,2,3]
plt.scatter(pvalues, coordinates)
plt.title('Plot 2')
plt.show()
```

3 Genome assembly and read mapping

Introduction to Genome Assembly and State-of-the-Art Algorithms

Genome assembly, a fundamental task in bioinformatics and genomics, involves the reconstruction of a complete genome sequence from fragmented DNA sequences. This intricate process is essential for understanding the genetic makeup of organisms, identifying genetic variations, and comprehending the molecular basis of various biological phenomena. As DNA sequencing technologies have advanced, producing enormous amounts of short DNA fragments, the field of genome assembly has witnessed remarkable progress in algorithmic development and computational techniques.

In the early days of genomics, the Sanger sequencing method generated relatively longer reads, simplifying the assembly process. However, with the advent of high-throughput sequencing technologies such as Illumina, which produce shorter reads, genome assembly became more challenging due to the complexities associated with reconstructing the original genome sequence accurately from these fragmented pieces. To address this challenge, a myriad of algorithms have been developed, each with distinct strategies, advantages, and limitations.

State-of-the-art genome assembly algorithms employ various innovative approaches to tackle the complexities of assembling genomes from short DNA reads. One prominent approach is the "de Bruijn graph" method, which involves breaking down the reads into smaller k-mers (subsequences of length k) and constructing a graph where nodes represent k-mers and edges depict their overlapping relationships. This graph-based representation enables the detection of overlaps and the identification of potential paths through which k-mers can be connected to reconstruct longer sequences.

Another cutting-edge technique involves the use of "overlap-layout-consensus" (OLC) approaches. In this strategy, reads are aligned to one another to identify overlapping regions, creating a layout of their relative positions. By establishing a consensus sequence from these overlaps, the algorithm constructs longer contiguous sequences. OLC methods are particularly advantageous for assembling longer reads generated by technologies like PacBio and Oxford Nanopore.

Additionally, hybrid assembly approaches combine data from different sequencing technologies to capitalize on their respective strengths. For example, short reads from Illumina can be used to correct errors in long reads from PacBio or Nanopore, resulting in highly accurate and contiguous assemblies.

Recent advancements in machine learning and artificial intelligence have also influenced genome assembly algorithms. Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been applied to improve error correction, scaffolding, and other assembly steps, enhancing the accuracy and efficiency of the overall process.

In conclusion, genome assembly stands at a fascinating crossroads of biology, computer science, and data analysis. State-of-the-art algorithms continue to evolve, leveraging innovative strategies and leveraging advancements in sequencing technologies and computational methods. These algorithms enable researchers to unlock the mysteries of genetics, uncovering insights into the structure and function of genomes with increasing accuracy and speed.

Sequencing technologies

Graph-based assembly

Dealing with repeats

Assessing the quality of an asssembly

Resequencing

4 Finding genes in bacteria ★

Some background

Do it yourself:

5 Pairwise alignment

An optimal alignment

- What is an alignment?
- What do we want it to represent?
- Scoring an alignment

A recursive solution

Global pairwise alignment

If i give you the following three optimal alignments: n-1, m-1; n, m-1; n-1, m can you the. Give me the optimal alignment of n,m?

The Needleman-Wunsch algorithm is a dynamic programming algorithm commonly used for global sequence alignment in bioinformatics and computational biology. Developed by Saul B. Needleman and Christian D. Wunsch in 1970, this algorithm is fundamental for comparing two sequences, such as DNA, RNA, or protein sequences, to identify similarities, differences, and evolutionary relationships between them.

Problem Statement: The primary objective of sequence alignment is to identify the optimal alignment between two sequences by inserting gaps (representing insertions or deletions) to maximize the similarity score while adhering to predefined scoring rules.

Algorithm Overview: The Needleman-Wunsch algorithm builds a dynamic programming matrix to efficiently calculate the optimal alignment score and traceback through the matrix to retrieve the optimal alignment itself. It employs a recurrence relation to calculate the score of aligning prefixes of the two sequences and incrementally fills the matrix based on these calculations.

Scoring Scheme: The algorithm uses a scoring scheme to assign scores to matches, mismatches, and gaps. Typically, a positive score is assigned to matches, a negative score to mismatches, and negative penalties for gap openings and gap extensions. The scoring scheme reflects the biological context and influences the alignment results.

Dynamic Programming Matrix: The dynamic programming matrix is constructed with dimensions M x N, where M is the length of the first sequence and N is the length of the second sequence. Each cell in the matrix represents the score of aligning the prefixes of the two sequences up to that point.

Initialization: The first row and the first column of the matrix are initialized based on the gap penalties. The first row corresponds to aligning the first sequence with gaps in the second sequence, and the first column corresponds to aligning the second sequence with gaps in the first sequence.

Recurrence Relation: Starting from the second row and the second column, each cell (i, j) in the matrix is calculated as the maximum of three values:

The cell above (i-1, j) plus the gap penalty (gap extension). The cell to the left (i, j-1) plus the gap penalty (gap extension). The diagonal cell (i-1, j-1) plus the match/mismatch score based on the characters in the sequences at positions i and j. Traceback: After the matrix is filled, the optimal alignment can be reconstructed by backtracking from the bottom-right corner (i.e., the end of both sequences) to the top-left corner (i.e., the beginning of both sequences). At each step, the decision to move diagonally, up, or left is based on the values in the neighboring cells and the scoring rules.

Alignment Output: The traceback results in the aligned sequences with gaps, showcasing the optimal alignment between the input sequences. The alignment score is the value in the bottom-right corner of the matrix and represents the degree of similarity between the two sequences.

Complexity: The Needleman-Wunsch algorithm has a time and space complexity of O(M * N), where M and N are the lengths of the input sequences. This complexity makes it suitable for relatively short sequences but can become impractical for very long sequences.

In summary, the Needleman-Wunsch algorithm is a foundational tool for performing global sequence alignment, allowing researchers to compare and analyze biological sequences to uncover evolutionary relationships, genetic variations, and functional similarities between genes, proteins, and other biological molecules.

Local pairwise alignment

Alignment significance

Do it yourself:

6 Protein substitution matrices

Scoring protein alignments

PAM matrices

The PAM (Point Accepted Mutation) matrices are a family of substitution matrices used in bioinformatics for comparing protein sequences. PAM matrices are derived from an evolutionary model that assumes a specific rate of amino acid substitution over time. The number in the PAM matrix represents the fraction of accepted mutations (amino acid substitutions) at a specific position over a certain evolutionary distance. Here's how to compute a PAM120 substitution matrix:

Collect a Set of Homologous Sequences: Gather a set of protein sequences that are believed to be homologous or evolutionarily related. This set of sequences will serve as the input for constructing the substitution matrix.

Calculate Pairwise Alignments: Align all pairs of sequences in the collected set using a suitable pairwise sequence alignment algorithm, such as Needleman-Wunsch or Smith-Waterman. This generates an alignment for each sequence pair, which helps identify positions where substitutions have occurred.

Calculate Mutation Frequencies: Count the number of times each pair of amino acids has been substituted for each other across all the alignments. This provides the raw data for calculating the substitution probabilities.

Calculate Fraction of Accepted Mutations (PAM1): For each amino acid pair, calculate the fraction of accepted mutations, which is the number of observed substitutions divided by the total number of aligned positions. PAM1 is the initial estimate of substitution probabilities.

Estimate PAM Matrix for a Specific Evolutionary Distance: The PAM120 matrix represents amino acid substitutions over an estimated evolutionary distance of approximately 120 PAM units (PAM120). To compute this matrix, you need to perform a series of calculations using the PAM1 matrix.

Iterative Calculation: PAM matrices are constructed through iterations. The general idea is to extrapolate from PAM1 to PAM120 through a series of transformations. Each PAM

matrix is derived from the previous one using a formula that relates the substitution probability in the current matrix to the substitution probabilities in the previous matrix.

Adjust for Total Probability: In each iteration, the substitution probabilities are adjusted to ensure that the total probability of amino acid substitution remains constant. This helps maintain the expected properties of a substitution matrix.

Normalization and Scaling: Normalize the calculated substitution probabilities to ensure they sum up to a certain value (often 1). Scale the probabilities by factors that depend on the number of substitutions and the evolutionary distance.

Rounding and Scaling for Matrix Values: Round the calculated probabilities to integers, ensuring that they fit into the matrix's integer-based format. Scale the values by a constant factor to control the scoring range.

Matrix Format: Present the calculated values in a matrix format, where rows and columns correspond to the 20 standard amino acids. The values in the matrix indicate the substitution scores between the respective amino acids.

It's important to note that computing PAM matrices requires a deep understanding of evolutionary models, bioinformatics algorithms, and statistical concepts. PAM matrices are typically provided as precomputed matrices for various evolutionary distances, and their calculation involves complex mathematics. Therefore, most bioinformatics applications use precomputed PAM matrices rather than attempting to compute them from scratch.

BLOSUM matrices

The BLOSUM (BLOcks SUbstitution Matrix) series of substitution matrices are widely used in bioinformatics for comparing protein sequences. BLOSUM matrices are derived from a set of aligned protein sequences and provide substitution scores that reflect the observed frequencies of substitutions between different amino acids. Here's how to compute a BLOSUM62 substitution matrix:

Collect a Sequence Database: Gather a diverse and representative set of protein sequences that are homologous or related to each other. This database will be used to calculate the substitution scores.

Create Sequence Pairs: Generate pairs of sequences from the collected database. These sequence pairs should be aligned to each other to identify the positions where substitutions occur.

Calculate Substitution Frequencies: Count the number of times each pair of amino acids is observed to be substituted for each other in the aligned sequences. This forms the basis for calculating substitution probabilities.

Calculate Substitution Probabilities: For each pair of amino acids, calculate the probability of substitution by dividing the observed substitution frequency by the total number of aligned pairs. This gives an idea of how likely one amino acid is to be substituted for another.

Calculate Log-Odds Ratios: Convert the substitution probabilities into log-odds ratios by taking the logarithm (usually base 2) of the probability of substitution divided by the background frequency of observing those amino acids independently.

Normalization: Normalize the log-odds ratios to ensure that the values are centered around zero and suitable for scoring alignments. This normalization step helps in preventing very high or very low scores.

Adjustment for Diagonal Elements: The diagonal elements of the matrix (representing substitution of an amino acid with itself) are usually set to a fixed value, typically a negative value, to discourage mismatches.

Rounding and Scaling: Round the log-odds ratios to integers and scale them by a constant factor to ensure that the substitution scores are suitable for the scoring range desired.

Matrix Format: Present the computed values in a matrix format, where rows and columns correspond to the 20 standard amino acids. The values in the matrix indicate the substitution scores between the respective amino acids.

BLOSUM62 is specifically derived from a sequence identity of around 62% and is suited for comparing protein sequences that share this level of identity. It is widely used for aligning moderately divergent protein sequences. Different BLOSUM matrices with different sequence identity cutoffs (e.g., BLOSUM45, BLOSUM80) are available to accommodate sequences with varying degrees of similarity.

It's important to note that BLOSUM matrices are typically precomputed and provided as standard matrices for various sequence comparison tools. If you're looking to generate a BLOSUM matrix from scratch, it involves significant computational and statistical analysis, and specialized software tools are often used for this purpose.

6.1PAM vs BLOSUM

BLOSUM (BLOcks SUbstitution Matrix) and PAM (Point Accepted Mutation) are both families of substitution matrices used in bioinformatics for comparing protein sequences. They are designed to quantify the likelihood of amino acid substitutions between sequences and are fundamental components in sequence alignment algorithms. Each family has its own advantages and limitations. Here's a comparison of the pros and cons of BLOSUM and PAM matrices:

BLOSUM (BLOcks SUbstitution Matrix):

Pros:

Empirical Nature: BLOSUM matrices are derived from a specific set of aligned protein sequences, often representing more recent evolutionary events. As a result, they capture more recent divergence and are suited for comparing moderately similar sequences. Variability: Different BLOSUM matrices are available (e.g., BLOSUM30, BLOSUM62, BLOSUM80), catering to varying degrees of sequence identity. This allows users to choose the matrix that best fits the similarity level of their sequences. Specificity: BLOSUM matrices can be used for aligning sequences with a wide range of similarity levels, making them versatile for various types of sequence comparisons. Scoring Range: BLOSUM matrices often result in a wider scoring range, providing more detailed information about sequence similarities and differences. Cons:

Limited Evolutionary Distance: BLOSUM matrices are less suitable for highly divergent sequences or sequences separated by long evolutionary distances. Pre-computed Nature: BLOSUM matrices are precomputed and can't be easily recalculated with custom data. This limits their adaptability to specific datasets. Homologous Sequences Required: The quality of BLOSUM matrices depends on the quality and diversity of the input sequences used to derive them. Alignment Quality Sensitivity: BLOSUM matrices are sensitive to the quality of the initial sequence alignment used to derive them. PAM (Point Accepted Mutation):

Pros:

Evolutionary Model: PAM matrices are derived from an evolutionary model, which attempts to capture long-term evolutionary processes. They can be used for comparing sequences with a broader range of evolutionary distances. Flexibility: PAM matrices can be theoretically recalculated for custom datasets, although it requires significant computational effort and expertise. Range of Matrices: Like BLOSUM, different PAM matrices (e.g., PAM30, PAM120, PAM250) are available for varying degrees of sequence divergence. Long-Distance Comparisons: PAM matrices are more suitable for comparing highly divergent sequences due to their focus on longer-term evolutionary events. Cons:

Dependence on Model Assumptions: PAM matrices rely on the assumption of a specific evolutionary model, which may not perfectly represent all evolutionary processes. Complexity: Constructing and recalculating PAM matrices require significant computational resources and expertise. Sensitivity to Input Sequences: Like BLOSUM matrices, the quality of PAM matrices depends on the quality and diversity of the input sequences. Non-Contemporary Alignments: PAM matrices are designed to capture historical substitutions, which can be a limitation when comparing sequences with recent divergence. In summary, both BLOSUM and PAM matrices offer valuable tools for comparing protein sequences. The choice between them depends on the specific characteristics of the sequences being aligned, the desired degree of sensitivity, and the computational resources available. BLOSUM matrices are preferred for more recent evolutionary events and

moderately similar sequences, while PAM matrices are better suited for longer-term evolutionary processes and highly divergent sequences. # Data base searching

BLAST

> BLAST is the 12'th most cited scientific paper - not just in bioinformatics but in all of science!.

Significance of search hit

Do it yourself:

7 Multiple alignment

How hard can it be?

Heuristic approaches

> ClustalW is the 10'th most cited scientific paper - not just in bioinformatics but in all of science!

Do it yourself:

8 Models of DNA evolution

Jukes-Cantor model

Kimura two-parameter model

GTRM

9 Clustering of sequences

UPGMA

Neighbor-joining

10 Phyogenetics

Likelihood of a tree

11 Hidden Markov models

What it is

Forward algorithm

Posterior decoding

Training / parameter estimation

12 Neural networks

13 Sequence annotation

What it is

14 RNA structure

What it is

References