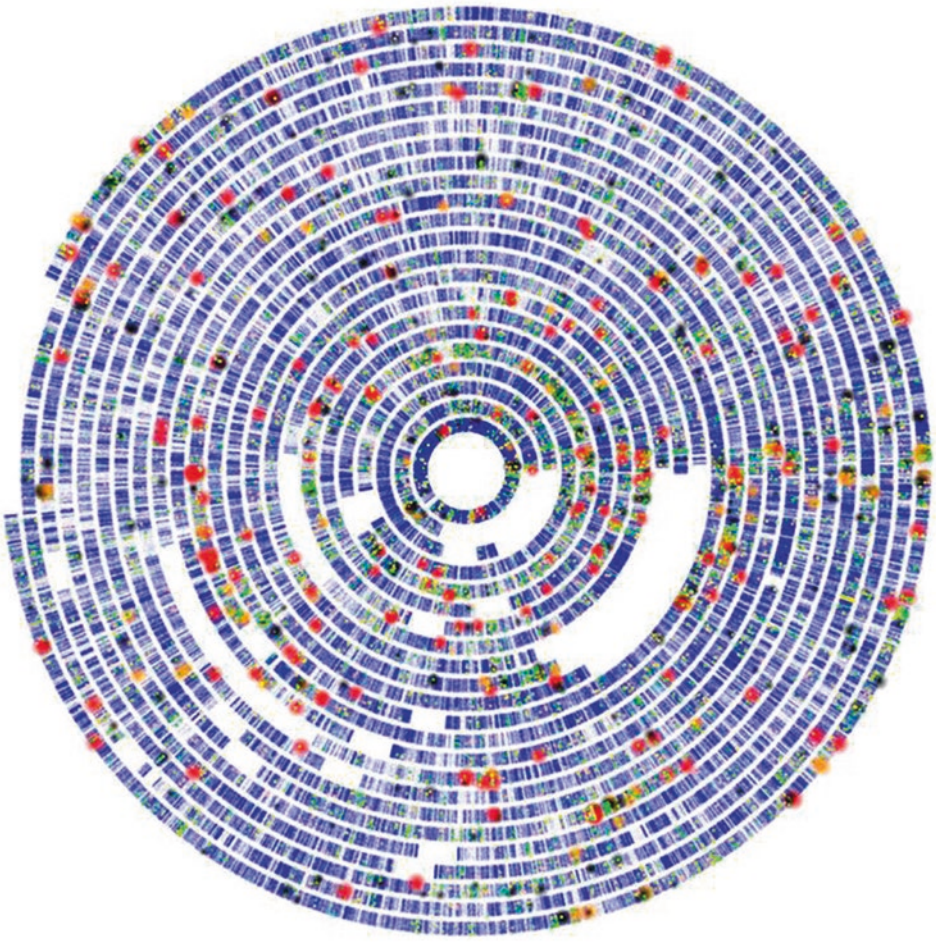
Thomas Dandekar
Meik Kunz

Bioinformatics

An Introductory Textbook

Springer

Bioinformatics



No black and white: Shown are the fascinating shades of individuality. In this artistic representation, all the variants of a healthy human being (NIH assembly identifier: NA12878) are displayed. They are organized on several circles, representing the different chromosomes, according to their position on the chromosome. The size and color of the variants were chosen according to the severity of the impact on the function of the genome. For example, you can see the many gray variants that do not fall on any gene and are therefore difficult to classify. This contrasts with the black and dark variants, which cause a severe defect in the affected genes. This shows how a considerable number of gene defects can be found even in healthy people as they are compensated by the healthy gene copy from the other parent.

Thomas Dandekar • Meik Kunz

Bioinformatics

An Introductory Textbook

Springer

Thomas Dandekar
Department of Bioinformatics
University of Würzburg
Würzburg, Germany

Meik Kunz
Chair of Medical Informatics
Friedrich-Alexander University
Erlangen-Nürnberg, Germany

ISBN 978-3-662-65035-6 ISBN 978-3-662-65036-3 (eBook)
https://doi.org/10.1007/978-3-662-65036-3

This Springer imprint is published by the registered company Springer-Verlag GmbH, DE, part of Springer Nature.
The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

Contents

Part I

How Does Bioinformatics Work?

Access

We are searching the key to understand life – this is how bioinformatics is oriented nowadays! It has evolved from data processing, just the assistant and auxiliary science for large amounts of data, to now establish *quantitative* theoretical biology. For the first time, theories about something as complex as living beings no longer remain pure theory, but are directly verifiable and measurable, and have already led to remarkable results and progress – from drugs against cancer and HIV to new insights, for example into the exciting question of why our cells and we age.

Nevertheless, my main motivation for studying medicine and later becoming a bioinformatician was not so much the prospect of ploughing through large amounts of data, but the fascination that biology has always had for people, the eternal questions about the key to the language of life, about the "water of life" that heals everything. I wanted to recognize and understand what holds us together in our innermost self, that is, how our consciousness and our brain function. Tracing these great questions is precisely the purpose of this book. Because today's bioinformatics is doing this to an increasing extent, and because one can also start from very small, simple examples, we will begin with these. We provide case-based examples for each chapter and a tutorial in the appendix for you to play with and discover for yourself. The new English edition 2021 brings everything up to date and adds further important aspects.

The unbelievable has happened silently: Whereas before the computer was just a stupid data storage device, new insights into life and the world and ourselves are now emerging in simulations. This is only possible because life itself is not dead and is permeated by numerous recognition processes. These are, for example, key-lock relationships between molecules, but also memory and molecular languages at all levels of life. We want to explore this in more detail here, first looking at the "how" of bioinformatics, in order to then better understand in Part II why bioinformatics is so successful right now – similar to

theoretical physics in the first half of the last century. This will also prepare us to explore the fascination of information processing in living beings and its reflection in the computer model (Part III), whether we want to better fight infections, understand cancer, or even understand ourselves.

Short Instructions for Usage of the Book

A classical textbook should (i) teach you the practice of bioinformatics and (ii) provide accurate definitions. For these two points, we have (i) prepared not only exercises in each chapter, but also tutorials for the most important software examples along with tips for use, and (ii) included a number of definitions in the glossary so that important terms are defined and explained.

Nevertheless, the book here is deliberately not a classical textbook. We want to convey joy and interest in bioinformatics. You can and are welcome to read the examples and chapters at your leisure and then, if you are interested in certain analyses in more detail, to practice them, work through the questions, look at the tutorials and do everything in more detail. Systematically, all current areas of bioinformatics are presented in a broad overview, and each end of chapter briefly summarizes the presented area again in a conclusion. We can only provide a stimulating introduction here. Without practicing and working through several examples for each of the software, it is not possible to gain sufficient experience for your own analyses. A sound knowledge of biology is also important, since you should be able to critically examine the program outputs with your knowledge. A number of suggested books on molecular biology but also on the national research data and medical informatics initiative are listed in the chapters. For students who enjoy programming, appropriate references for further reading are given in the introduction to the tutorials. Since bioinformatics lives on databases and software, we have summarized databases and programs and their basic use in the chapters and in the appendix.

Sequence Analysis: Deciphering the Language of Life

1

Abstract

Sequence analysis is a central tool of bioinformatics with relevant databases (NCBI, GenBank, Swiss-Prot) and software to detect sequence similarity (BLAST) and domain databases (Pfam, SMART). Crucial is the ability to know and use such software on the web, the tutorials and exercises encourage this. Programming sequence comparison software and databases only makes sense if it enables a better analysis of the biological question, in particular for large-scale analysis – in all other cases, it is better to use the numerous software that already exist, the internet is only a mouse click away.

Bioinformatics requires data on living organisms, processes them and then designs a corresponding model of the living process that is thereby mapped. A good simple example is when a polymerase chain reaction (*PCR*) is used to detect a virus in the blood. Polymerases copy DNA (*deoxyribonucleic* acid) and were originally derived from bacteria. Hereby they also duplicate their genetic information. PCR is a modern method of molecular biology. Using such a chain reaction, so much of a molecule (if, for example, there is only one virus molecule in the blood) is produced by constant doubling of the molecules with the help of polymerase that it can be easily detected in the laboratory and, above all, the sequence can be read.

Nowadays, this can be deciphered quite easily by a sequencing machine. However, this initially leaves us with a salad of letters that lists the nucleotides, i.e. the genetic material, of the virus in sequence, such as tgtcaacata ... (Fig. 1.1).

T. Dandekar, M. Kunz, *Bioinformatics*,
https://doi.org/10.1007/978-3-662-65036-3_1

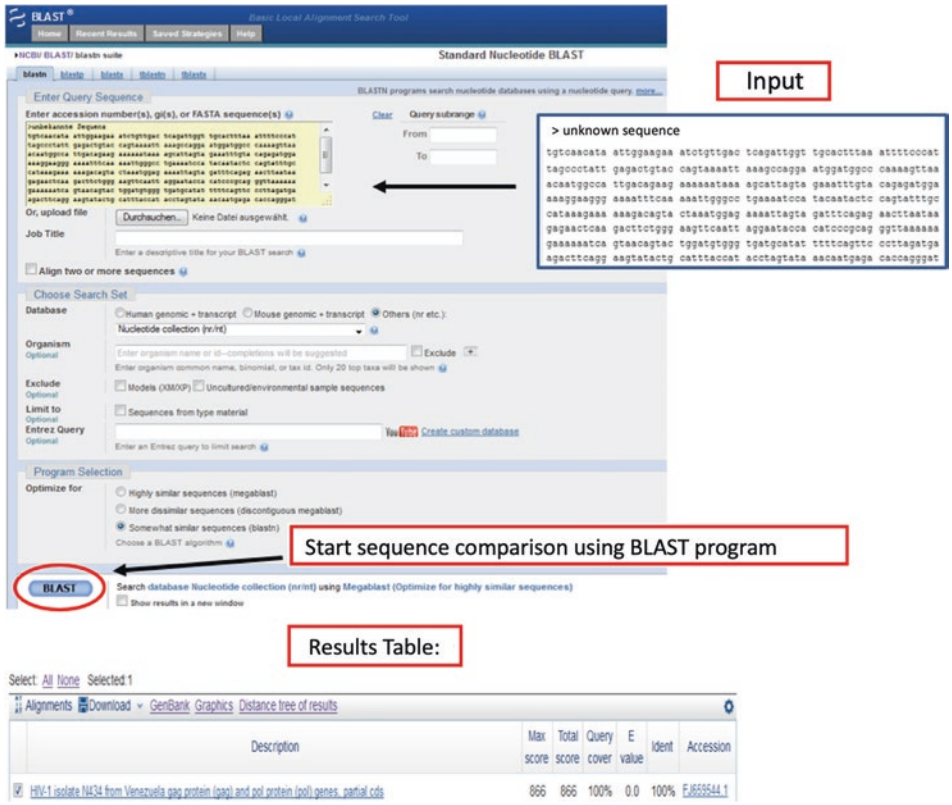


Fig. 1.1 Sequence analysis allows identification of HIV virus sequences. HIV sequence identification using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Shown is the sequence comparison of an initially unknown sequence against a database using the program BLAST. The result line indicates that the unknown sequence is an HIV-1 N434 retrovirus strain from Venezuela (result line: Venezuela gag coat protein and pol polymerase protein; the result link then leads to the detailed sequence comparison)

Collect, Compare and Understand Data In order to now know which virus we have in front of us (in practice, usually even much more precisely, namely which virus strain), we have to let the computer identify this sequence.

Collect Data This is particularly easy if you have created a database of virus sequences. You already know their sequence because you have sequenced them before. As an example, let us consider HIV, the human *immunodeficiency* virus. With the help of the database, it is easy to find out whether the sequence found by PCR for a virus in the blood matches one of the entries in the database. Databases are fundamental in bioinformatics. They store all the information and can then be used for further investigations.

Analyze and Compare Data

So this is how you do a sequence comparison (also called sequence analysis). You look to see which sequence in the **database** is most similar to the new sequence. This can be done over the entire length of the sequence, i.e. globally. However, because a virus can be relatively strange and one would then usually like to know whether it is not at least similar in sections, one typically performs a section-by-section local comparison, which thereby yields the most similar sequence section (Fig. 1.1). But in order for the computer to do anything at all, you have to tell it what to do down to the last detail, until it finally presents a result of the computation. All the instructions for this, e.g. to perform such a comparison up to the final result, are together a program. In the past, **programs** were written using instructions that the machine understood particularly well. But these could only be very short, because they were written in machine language, which essentially contained simple register instructions (clear 1 bit, write, move or check). Today, however, a richer language is used that contains far more complicated instructions, which is therefore called a higher programming language (e.g. Perl, Java, Python, C++ or R, currently the most popular programming languages in bioinformatics).

Let us return to our sequence example: What do we see as a result in Fig. 1.2? This is a so-called *Basic Local Alignment*, the corresponding tool in bioinformatics is called BLAST, for *Basic Local Alignment Search Tool* (Altschul et al. 1990), where the result indicates a veritable diagnosis for the patient.

The sequence comparison shows that it is an HIV strain from Venezuela. It becomes clear that one can actually make a diagnosis (HIV infection, probably acquired in South

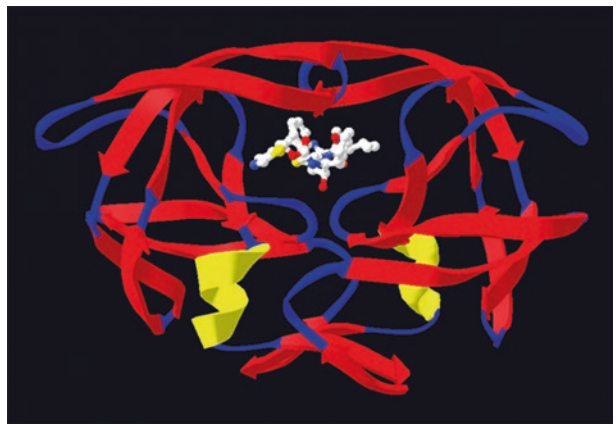


Fig. 1.2 *Drug design*, example of HIV infection. The HI virus is blocked in its activities (dark molecule around the drug) by a drug (centre, white). Computer representation of the three-dimensional structure of the HIV-1 protease (molecular structure consisting of leaflets [red], loop regions [blue] and helices [yellow]) and its inhibitor ritonavir (shown as a sphere and edge model). The aim of such bioinformatic *drug designs* is to design a suitable therapy on the computer, in this case, for example, the inhibition of the protease for the treatment of an HIV-1 infection, so that the virus can no longer produce new viral envelopes - its protease no longer functions

America) with this computer program, which only writes letters as optimally as possible among each other (hence sequence comparison or alignment). The decisive prerequisite for this is that one knows and understands the results correctly in their biological meaning - and this is precisely the work of the bioinformatician.

Understanding Data

Finally, there is a third area of work in bioinformatics: "understanding data". In addition to collecting data (databases) and comparing data (e.g. using BLAST), one ultimately wants to understand the data and use it appropriately, for example to develop new therapeutic approaches. This can happen, among other things, by integrating the data in a suitable **bioinformatics model** and then modelling it. This modelling can be a simulation, for example if I am looking for new drugs against HIV and want to destroy the sequence of the virus. Since the virus consists of nucleic acids, as we have already seen above, I can, for example, insert the wrong nucleotides into the virus and thus also destroy its polymerase (the copying enzyme with which the virus reproduces). A complex but highly successful modelling technique consists of reproducing the three-dimensional structure of this polymerase in the computer and then selecting from a database of molecules which best fits into the polymerase in such a way that it is blocked, i.e. the virus can no longer reproduce (Fig. 1.2 shows an example of this *drug design*). Such methods have been very successful with HIV in particular. There are now more than 20 drugs that target the virus with the wrong nucleotides, by inhibiting its nucleic acid or its enzymes. The result is remarkable, the combination therapy (*highly active* antiretroviral therapy; HAART, Antiretroviral Therapy Cohort Collaboration 2008) works so well that one has an almost normal life expectancy, while one can only withstand the viral infection for a few years without therapy (Hoog et al. 2008). This illustrates that bioinformatics can strongly support medicine for instance regarding therapy.

What would you actually have to pay special attention to if, for example, you now perform such sequence comparisons yourself? It is important to know that the BLAST search is not completely accurate (heuristic), but it delivers faster results than a 1:1 comparison over the entire sequence length against the database. Therefore, such hits are only credible if the probability of getting such a hit by chance is low enough. As a first rule of thumb you can remember: The *E-Value* (i.e. the expected value of a random hit) should be less than 1 in one million. This is then already a very convincing value. In borderline cases (random expectation value at 1 in 1000), you can also take the hit sequence and see if you can find the initial sequence again (called "reverse search" in technical jargon). If we keep in mind that this is a local search, then we also understand why we should search the whole hit length (given in the example, sequence similarity over the whole sequence length). But there are also BLAST results where only one subsequence in the protein has high similarity and the rest instead shows no similarity. In this case, the BLAST search turned up only one protein domain, the one with the highest similarity in the whole database. To determine the remaining parts of the sequence in terms of function as well, you then need to use only those domains that do not yet have database hits again, without the first sequence part

for the search. In this way, you can match domain by domain in the protein with a new BLAST search each time for the sequence portion that has not yet been matched by the search. Finally, in difficult cases, the BLAST search may only reveal a similarity to a database entry that has no clear function. In this case (protein sequence), you can use the "position-specific iterative BLAST", or Psi-BLAST for short, which then searches with all the still unrecognized sequences at the same time (a so-called "profile") until it lands a hit to which a sequence can be assigned. This almost always works, but may take several repetitions. You should also only continue searching with Psi-BLAST if something changes in the repeat search, otherwise the search is "converged" in vain.

However, the drug search shown in Fig. 1.2 is a somewhat involved process, requiring many intermediate results to be obtained and calculations and comparisons to be made. What can be done, on the other hand, is to perform direct databases that provide additional secondary information besides the primary sequence information. These are also called secondary databases. An example would be to search for the HIV protease in the protein database PDB (https://www.rcsb.org/pdb/home/home.do). In addition to the protein sequence, this database also holds the coordinates of the protein's three-dimensional structure, as well as other details about its structure and function. There is a great deal of further information available on the HIV structure in particular, including information on the *drug design*.

1.1 How Do I Start My Bioinformatics Analysis? Useful Links and Tools

Generally speaking, we first look at the function of the molecule we want to bioinformatically determine by comparing it directly to a database. The best known example is the direct sequence comparison with BLAST, which we have already discussed in detail. The next step is to use other databases or programs for analyses and comparisons to obtain additional information. A simple example is to search for secondary data, and our first example of this was the protein database. As a primary database, it contains the three-dimensional coordinates of protein structures, but it also contains a lot of secondary data about these proteins where this structure determination was successful. As a third step, we can finally follow up with detailed analyses.

In the following, useful supporting sites for these steps are briefly presented. The BioNumbers database describes number relationships in biology (https://bionumbers.hms.harvard.edu). This was established at Harvard University by students who first calculated these biological problems and then made these numbers available to the interested reader.

Unfortunately, most bioinformatics websites are in English, including this book. This is due to the fact that the Anglo-Americans were simply faster with many initial developments than German bioinformatics. In addition, English is now the language of science, and the creator of a bioinformatics website would like everyone to be able to use this site.

Already Prepared Results: "BioNumbers"

▶ https://bionumbers.hms.harvard.edu/

So here you can find out how different sizes and numbers are related in biology. Just look it up and learn about the exciting world of sizes and numbers in different organisms and diseases, but also in humans.

For a better understanding, we would like to show a simple *screenshot* of a list of useful biological quantities and numbers from the BioNumbers database (Fig. 1.3). It is best to simply look at it yourself and be amazed at the interesting correlations and differences.

MEDLINE as a Large Online Library

One of the main problems in all bioinformatics work is to get a quick overview of the knowledge that exists about the object of study. This is the only way to assess the accuracy

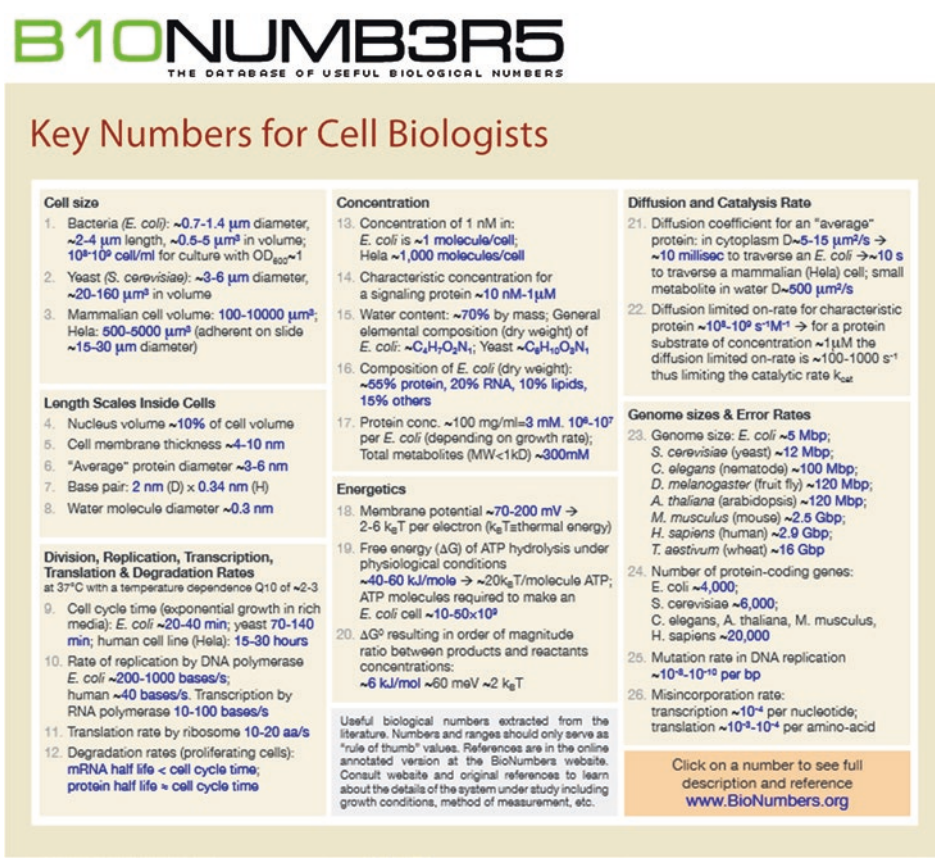


Fig. 1.3 Listing of useful biological quantities and numbers from the literature in the BioNumbers database (for details see text)

and also the value of your results. For this purpose MEDLINE, the online version of the library at the National Institute of Health, is an indispensable tool. A large, worldwide open library about medicine and biology:

- MEDLINE (or also PubMed)

- https://www.ncbi.nlm.nih.gov/pubmed

It is the online version of the library. Only here, in Betheds (near Washington), the Health Research Center of the United States of America, has it been possible to keep a sufficiently large staff of service scientists permanently on hand to ensure easy use of the web pages and to keep the data constantly up to date. This is a truly extraordinary achievement, which is precisely why it looks and feels child's play to use.

Here you can search for keywords ("HIV", "*sequence analysis*", "*aging*"), for authors ("Dandekar-T", "Kunz-M"), journals ("Nature", "Science"). For each article found, a table of contents will then appear, but also links to related articles (including search options). A steadily increasing number of articles also offer a directly readable full-text link ("*Open Access* ", even for current articles already more than 30%, for articles one to 2 years old it is now even the majority). It is possible for the experienced to search for an article much more precisely and with many more criteria ("*advanced search*"). It is helpful to have a look at the PubMed tutorials or our tutorial in the appendix. In addition, PubMed also provides important textbooks online and a variety of other resources.

How Do I Get the Sequence to My Molecule?

Many bioinformatics studies start with the sequence of a molecule and analyze it. Interestingly, this important starting information, i.e. what sequence the molecule I am interested in has, is already known for many millions of entries. This is especially true for important organisms such as humans, the bacterium *Escherichia coli (E. coli)*, plants such as *Arabidopsis*, mice, the worm *Caenorhabditis elegans (C. elegans)*, and the fruit fly *Drosophila melanogaster*. To check if my sequence for this protein or term is already known, look it up at NCBI in particular. If it is known, the sequence for DNA, RNA (option "*nucleotide*" or "*gene*") or proteins (option "*protein*") *can* easily be found here, e.g. for "HIV" there are hundreds of thousands of entries:

- https://www.ncbi.nlm.nih.gov/protein/?term=hiv

One of the first offers from the long list of hits is an artificial sequence for the "TAR protein":

- https://www.ncbi.nlm.nih.gov/protein/AAX29205.1

The now mostly quite long header entry explains already existing information about the respective protein:

```
LOCUS          AAX29205    367 aa linear    SYN 29-MAR-2005
DEFINITION     TAR, partial [synthetic construct].
ACCESSION      AAX29205
VERSION        AAX29205    .1 GI:60653021
DBSOURC        Eaccession     AY892288.1
KEYWORDS       Human     ORF project
SOURCE         synthetic     construct
ORGANISM       synthetic     construct
```

... and so on. In particular, you can find information about the authors of the sequence, journal articles about it and the exact properties of the sequence, that is, from where to where, for example, the protein, the region and specific binding sites go:

```
Protein        1..>367
               /product="TAR"
Region         30  ..95
               /region_name="DSRM"
               /note="Double-stranded RNA binding motif. Binding is not
               sequence specific but is highly specific for double stranded
               RNA. Found in a variety of proteins including dsRNA depen-
               dent protein kinase PKR, RNA helicases, Drosophila staufen
               protein, E. coli RNase III; cd00048"
               /db_xref="CDD:238007"
Site           order(30,36..37,78..81,84)
               /site_type="other"
               /note="dsRNA binding site [nucleotide binding]"
               /db_xref="CDD:238007"
Region         159    ..222
               /region_name="DSRM"
               /note="Double-stranded RNA binding motif. Binding is not
               sequence specific but is highly specific for double
               stranded RNA. Found in a variety of proteins including
               dsRNA dependent protein kinase PKR, RNA helicases,
               Drosophila staufen protein, E. coli RNase III; cd00048"
               /db_xref="CDD:238007"
               Siteorder(159,165..166,208..211,214)
               /site_type="other"
               /note="dsRNA binding site [nucleotide binding]"
```

Finally, this is followed by the original sequence as determined by the authors and used in their research. In the example:

```
ORIGIN
          1  mseeeqgsgt ttgcglpsie qmlaanpgkt pisllqeygt rigktpvydl
             lkaegqahqp
```

```
     6  1nftfrvtvgd tsctgqgpsk kaakhkaaev alkhlkggsm lepaledsss
        fspldsslpe
    12  1dipvftaaaa atpvpsvvlt rsppmelqpp vspqsecnp vgalqelvvq
        kgwrlpeytv
   181  tqesgpahrk eftmtcrver fieigsgtsk klakrnaaak mllrvhtvpl
        dardgnevep
   241  dddhfsigvg srldglrnrg pgctwdslrn svgekilslr scslgslgal
        gpaccrvlse
  3011  seeqafhvs yldieelsls glcqclvels tqpatvchgs attreaarge
        aarralqylk
   361  imagskl
```

The NCBI site brings a lot more information for bioinformatics:

https://www.ncbi.nlm.nih.gov/guide/ ...

All resources	A detailed overview of molecular and literature analysis and data banks
Chemicals and bioassays	Bioinformatic analyses should eventually lead to new experiments to confirm the results; the necessary ingredients and measurement methods are collected here: Chemicals and biological measurement methods (*bioassays*)
Data and software	Here we find numerous databases and programs
DNA and RNA	Software and tools for the analysis of DNA and RNA
Domains and structures	Analysis of protein domains (small folding units) and large structures
Genes and expression	Analysis of the transcription of genes under different conditions
Genetics and medicine	Numerous genetic information
Genomes and maps	Useful *maps to* find your way around genomes
Homology	Similarity comparisons to proteins, but at the structural level. In particular, it is thus possible to calculate one's own protein structure by pointing out a similar three-dimensional structure
Literature	In addition to MEDLINE (see above), there are many articles that can be found on the site and read online, as well as important textbooks
Proteins	General analyses of protein sequence, structure and function. In particular, the protein domains, i.e. the functional building units in the protein, are also examined in more detail
Sequence analysis	Other programs besides BLAST that examine the sequence of a protein or a nucleic acid
Taxonomy	Classification of a sequence in a catalogue of all species. Many of the results are presented as phylogenetic trees
Training and tutorials	Highly recommended for a first start, see: https://www.ncbi.nlm.nih.gov/guide/training-tutorials/ Especially the BLAST search and the *taxonomy* are explained in a very nice beginner tutorial
Variation	How do I do justice to biodiversity and variety?

In addition to the NCBI site, which is certainly the best known bioinformatics site, there are also good introductory sites at the European Bioinformatics Institute (EBI). These are especially helpful for those people who also like programming modules and are looking for information at an advanced level:

- https://www.ebi.ac.uk

For example:

- https://www.ebi.ac.uk/services

"*We maintain the world's most comprehensive range of freely available and up-to-date molecular databases*." This refers to the wealth of data that the EBI site offers. The difference to the NCBI website is that it is easier to download the entire data of the database and not only to perform individual queries via the web interface.

It is also important that the EMBL database is located here, which provides comparably detailed sequence information as GenBank at the NIH. However, there are small differences in the preferences and the offer, but also in the preparation of the entries. In addition, there is somewhat more and somewhat faster information on new sequences identified in Europe (NCBI is more detailed and faster for American sequences).

Other important sites can be found at the Swiss Bioinformatics Institute (see next chapter) and at the Japanese gene bank DDBJ (DNA Data Bank of Japan).

- https://www.ddbj.nig.ac.jp

Again, there is a daily comparison with the EMBL and NCBI databases in order to keep "all known" sequences available. This time, however, this is done from the Japanese point of view; it is precisely the sequences from Japan that are particularly complete and quickly recorded here.

Finally, reference should also be made to the new German National Research Data Infrastructure, in which targeted digitisation and infrastructure is being promoted in numerous areas.

- https://www.nfdi.de, https://www.nfdi.de/konsortien-2

For biology, for example, DataPlant (plant databases), the German Human Genome-Phenome Archive, NFDI4BioDiversity and NFDI4microbiota. This is also where very useful data for bioinformatics analysis is concentrated and made available as an infrastructure for all.

- https://nfdi4microbiota.de (Dandekar is an affiliate).

In addition, within the framework of the Medical Informatics Initiative of the Federal Ministry of Education and Research, there are several Germany-wide consortia to which university hospitals and other partners (research institutes, universities, companies) have joined forces.

- https://www.medizininformatik-initiative.de/de

For example, ten universities and university hospitals, two universities and one industrial partner are working together in the MIRACUM consortium (Medical Informatics in Research and Care in University Medicine) to establish an IT infrastructure for data from research and patient care (data integration centres) and to make it usable for research projects in the long term, for example for the development of predictive models and precision medicine.

- https://www.miracum.org/ (consortium leader Medical Informatics FAU Erlangen-Nürnberg, Kunz is a partner).

1.2 Protein Analysis Is Easy with the Right Tool

An important special case is the analysis of proteins. Many experiments in molecular biology focus on this particularly important type of molecule. Typically, general properties are first determined by experiments, such as certain binding sites, the weight of the protein, appearance, cofactors or catalytic properties. This is followed by detailed biochemical analyses. The Swiss Bioinformatics Institute has compiled a detailed software package for these numerous ways of analysing proteins. The site is again in English because such analyses are carried out here from all over the world, namely with regard to the properties of the protein sequence (secondary structure, amino acid composition and properties, antigenicity, etc.) as well as the protein structure, including the properties of the independent folding units in the protein, the protein domains.

Analysis with BLAST

A good first step is the already mentioned BLAST. This allows a protein sequence (blastp) to be compared for similar entries in a database, and also identifies conserved domains and motifs, such as catalytic and active sites.

In addition, there are more precise and specific tools, which are presented below.

Entry Page on the Web: ExPASy (https://www.expasy.org)

The Swiss Bioinformatics Institute had initially (1990s) built up the Swiss-Prot database under the direction of Amos Bairoch. It was particularly carefully maintained and still has a very high degree of correctness and correction of entries, even though it has now essentially been absorbed into the UniProt Knowledge base (UniProt KB):

▶ https://web.expasy.org/docs/swiss-prot_guideline.html

takes the interested person to this link. As explained on the page, there are also detailed comments on the sequence here. These so-called "header entries" provide a wealth of information about protein sequences, followed by the actual sequence.

How Do I Quickly Analyze Protein Data?

The ExPASy site brings expert help to get started with protein analysis. "*Proteomics*" means the analysis of large amounts ("*omics*") *of* protein data.

▶ https://www.expasy.org/proteomics

In addition to various databases, you can also find a lot of bioinformatics information here:

Proteomics	Large-scale analyses of proteins
Protein sequences and identification	Identification of proteins by sequence
Mass spectrometry and 2-DE data	Identification of peptides found in mass spectroscopy or protein spots found in 2D gel. Evaluation software and databases for these steps
Protein characterisation and function	Domain analyses in particular
Families, patterns and profiles	Proteins with the same function form a family. In particular, always the same ("conserved") amino acids, patterns and position-specific frequencies of amino acids for these families are summarized here
Post-translational modification	After production at the ribosome, proteins are further modified, these are the post-translational modifications
Protein structure	Finding or calculating the three-dimensional protein structure. A fast homology prediction via the SWISS-MODEL server is also offered here
Protein-protein interaction	Predicting which protein interacts with which other protein
Similarity search/ alignment	There are also a number of alternatives to BLAST here. Multiple protein sequences can also be compared
Genomics	How are the associated genes related to the proteins they encode?
Structural bioinformatics	In particular, the properties of protein structures are determined, for example globular proteins are particularly soluble
Systems biology	A nice introductory page on system effects of proteins, for example protein signalling cascades and phosphatases to switch off such signals
Phylogeny/evolution	Proteins develop according to specific patterns; in particular, building units, the protein domains, are assembled to form new proteins
Population genetics	How are important proteins and protein properties distributed in a population? What are the different types?
Transcriptomics	How are protein and its coding mRNA related?

(*continued*)

(continued)

Biophysics	What are the biophysical properties (solubility, stability, helix content, etc.) of my protein?
Imaging	How can proteins be visualized and images analyzed?
IT infrastructure	Computer infrastructure, service
Drug design	Helping to create new drugs to specifically target a protein
Glycomics	How sugar residues further modify proteins. In particular, this is how cells recognise their cell neighbours, bacteria cling to glycoproteins. Sugar-binding proteins are called lectins

How Do I Identify Important Amino Acids for Protein Function?
The PROSITE page is particularly helpful for this.

▶ https://prosite.expasy.org

This examines an entered protein sequence to determine whether or not certain sequence motifs are preserved, for example signatures (hand-curated) or profiles (automatically calculated, consensus sequences, taking different sequences into account) that indicate a particular enzyme function.

This allows me to check whether my protein sequence is really an active enzyme (then all amino acids for catalysis are complete) or whether it only looks like one. If this happens in a genome sequence, this is termed a "pseudogene", a "false" gene regarding the enzyme function because important catalytic amino acids are missing and the enzyme therefore cannot function.

In addition, the independent folding units in the protein, the protein domains, are also examined to see whether they are present in the protein, e.g. whether all parts, i.e. domains, are present for a functional enzyme: at least one catalytic domain (50–150 amino acids) that carries out the enzymatic reaction. This is then often joined by numerous other types, e.g. DNA interaction if it is a transcription factor. Examples are:

- cofactor-binding domains (if the enzyme binds a cofactor),
- regulatory domains (for switching the enzyme on and off),
- interaction domains (with other proteins or to form dimers of two identical protein units for the enzyme, e.g. glutathione reductase only functions as a dimer, so needs an interaction domain for its function),
- structural domains (e.g., if it is a structural protein, like collagen).

How Can I Estimate the Protein Structure?
Structure prediction with homology modelling, for example by SWISS-MODEL, is helpful for this.

▶ https://swissmodel.expasy.org

SWISS-MODEL offers the possibility to predict the three-dimensional structure of the protein based on the sequence.

This is a relatively quick prediction, and the three-dimensional coordinates are then available for the user to download. However, it requires a protein with a known three-dimensional structure as a template in order to calculate how much the user's sequence differs from this in its three-dimensional structure. Whether a template can be found is determined by a special sequence comparison with the proteins in the SWISS-MODEL database.

SWISS-MODEL is a very solid, fast and often confirmed approach to determine a three-dimensional structure according to protein template. However, there are many other, often much more complex ways of calculating the protein structure (e.g. homology modelling with MODELLER):

- https://salilab.org/modeller/tutorial/

Since structures are not always available that can serve as a template, so-called *ab initio* and optimization algorithms calculate an approximate solution for the structure determination based on the sequence and the minimization of the free enthalpy. Prominent representatives here are neural networks, evolutionary algorithm or Monte Carlo simulation. One example is the QUARK server from the Zhang lab:

- https://zhanglab.ccmb.med.umich.edu/QUARK/

Marking of the Known Structural Parts in the Protein Sequence

For independent verification, we offer at the chair a labeling of the known three-dimensional structural domains to any sequence (the technical language says domain annotation, that is why our tool is called "AnDom"). This is a slightly different procedure and works for any sequence. It just looks to see if at least a small piece of the sequence is not similar to a known three-dimensional protein structure. Thus, it is completely independent of the ExPASy predictions and can check them. In general, independent databases and softwares from different authors and methods check each other. This allows to significantly increase the quality of the predictions, e.g. to collect all structure predictions (broad search) or to accept only those found by both websites (particularly validated predictions).

This then sometimes makes the predictions a bit tight. This happens when only short parts of the sequence have sufficient similarity to the structural databases that AnDom has. It can also happen that the protein structure is new, i.e. not similar enough to any known structure to allow prediction. Just as when using BLAST, very small random expectation values (1 in one million and lower probabilities) mean that the assignment using AnDom has been very successful in revealing a structure similarity. In contrast, a random similarity can be recognized by a high random hit rate (higher than 1 in 1000). It may even happen that such a small similarity is found several times even by a random sequence. In this case, the expected value is e.g. 4, if on average a random sequence would find four such hits in the AnDom structure database.

▶ https://andom.bioapps.biozentrum.uni-wuerzburg.de/index_new.html

Again, the HI virus from Fig. 1.1 will serve as an example here (Fig. 1.4). AnDom finds a protease domain in the protein sequence (top: b.50.1.1 according to the SCOP classification). The alignment is also shown (bottom), which once again shows the high degree of agreement between the search sequence (query) and the protease domain found (Sbjt = *subject*) (93% identical). Please also use our tutorial for further information.

Conclusion

- In this first chapter, you have already quite actively learned and practiced the most important technique in bioinformatics, namely sequence analysis, especially of protein sequences. Modern molecular biology generates sequences in abundance. The steady increase of databases (NCBI, GenBank, Swiss-Prot) allows one to quickly find out which previous sequences are close to this new sequence by sequence similarity (BLAST tool). Domain databases and analyses allow to dissect a protein into its folding units, each of which carries a specific molecular function. RNA and DNA sequences are also quickly assigned a function through sequence comparisons.

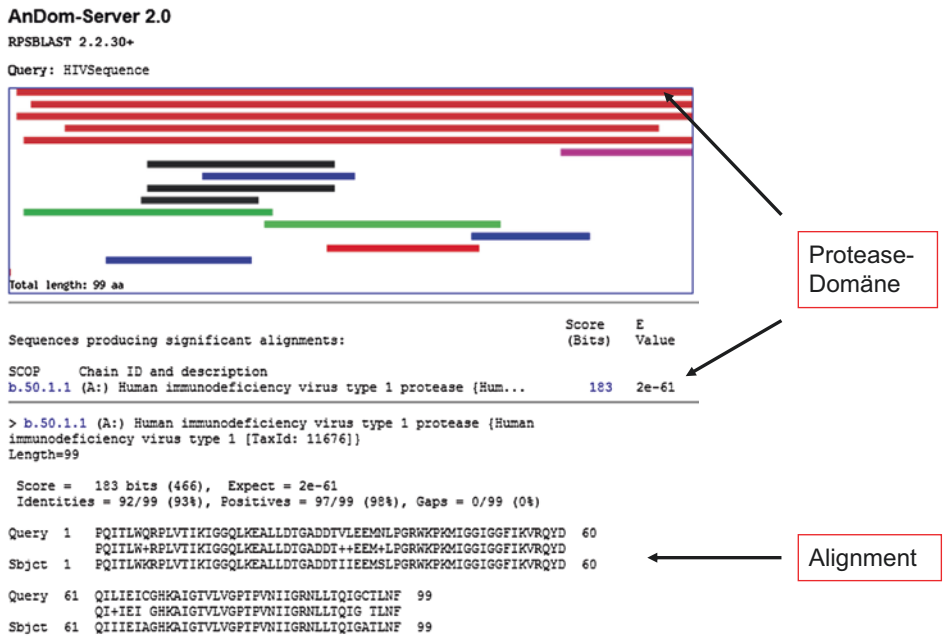


Fig. 1.4 Search with the AnDom software for protein domains for the HI virus (for details see text). The result shows a high similarity (E-Value 2e-61, 93% *identities*) with the human HIV-1 protease domain (SCOP-ID b.50.1.1) and the corresponding alignment (see text and tutorial)

- Undeniably, sequence analysis is currently the field of bioinformatics that is growing the fastest, producing the quickest results, and allowing initial insights into biology. Hence, in the later chapters, there is sequence analysis software that allows us to quickly trace partial results. It is crucial to be able to learn about this software on the web and practice the different setting options.
- The tutorials and exercises encourage you to do so. Results from different software programs check each other. If they all examine the same sequence, it is always about the same biology, and contradictions then indicate that something was overlooked in the function assignment and must be checked. Sound biological knowledge should critique the results, experiments or further data then corroborate the bioinformatic results. Programming sequence comparison software and databases is useful if this enables a better analysis of the biological question - in all other cases, it is better to use the numerous software that is already available. The internet is only a mouse click away. ◀

Outlook

In addition to protein sequence analysis (Chap. 1), RNA (Chap. 2) and DNA sequences (Chap. 3) are important for rapid bioinformatics analysis and description of important molecules of the cell. Next, one would like to understand how these important molecules of the living cell (DNA, RNA, and proteins) interact in networks. These bioinformatic analyses happen either in metabolic networks (Chap. 4) or signaling networks (Chap. 5). Since these are already the most important analysis techniques of current bioinformatics, we then offer an in-depth look at basic strategies of bioinformatics working methods in Part II and look at fascinating examples of current bioinformatics results and developments in Part III.

1.3 Exercises for Chap. 1

In the exercises, important parts of the book will be dealt with in more detail in order to consolidate and practise what you have learned. Tasks marked as examples serve as application tasks in which you are to work independently with the computer in order to become more familiar with bioinformatics. In addition, we have provided numerous tutorials in the appendix, which also support the material of the textbook and the exercises and should contribute to a better understanding.

We recommend that you briefly review the material from Chap. 1 at Chap. 6 using the exercises.

Task 1.1

(a) What is and does bioinformatics do (feel free to explain with an example)?
(b) There are three areas of bioinformatics, informatically speaking: Databases, Programs/Software, and Modeling/Simulations. Describe important differences between these areas.

Task 1.2

An important task of bioinformatics is the collection and management of data and the provision of helpful tools. Name and describe two databases containing information on, for example, genes and gene expression datasets.

Task 1.3

Example:

The MEDLINE database (also known as PubMed) is a large, worldwide open library about medicine and biology. Here you can find publications and sequences as well as a lot of other information and links. So PubMed is a good first entry site to use when starting a search. Familiarize yourself with the PubMed database (https://www.ncbi.nlm.nih.gov/pubmed) and find out about the *artificial* sequence for the "TAR *protein*". Hint: Search with "*synthetic*", all searches are in English after all; the search is only limited enough by keywords if only one sequence is found by the query. Only then can you clearly answer the following questions.

1. Which of the following statements about sequence length (amino acid = aa) is correct?
   A. The protein sequence is 267 aa long.
   B. The protein sequence is 367 aa long.
   C. The protein sequence is 276 aa long.
   D. The protein sequence is 376 aa long.
2. Which of the following statements about the title is correct?
   A. The sequence was filed under the title "Cloning of human full-length CDS in Creator (TM) recombinational vector system" in PubMed.
   B. The sequence has been filed under the title "Uploading of human full-length CDS" in PubMed.
   C. The sequence has been filed under the title "Uploading of recombinational vector system" in PubMed.
   D. The sequence has been filed under the title "Cloning of recombinational vector system" in PubMed.
3. Which of the following statements is correct?
   A. Hines et al. submitted the sequence to the journal *Biological Chemistry* and Molecular Pharmacology, Harvard Institute of Proteomics on 05-JAN-2015.
   B. Darwin et al. submitted the sequence to the journal *Biological Chemistry* and Molecular Pharmacology, Harvard Institute of Proteomics on 05-JAN-2005.
   C. Hines et al. submitted the sequence to the journal *Biological Chemistry* and Molecular Pharmacology, Harvard Institute of Proteomics on 05-MAR-2005.
   D. Hines et al. submitted the sequence to the journal *Biological Chemistry* and Molecular Pharmacology, Harvard Institute of Proteomics on 05-JAN-2005.

Task 1.4

Bioinformatics has taken off since the mid-1990s, when the first genome projects were successfully completed, because of its rapid sequence analyses. Sequence comparison (for

example with the BLAST software) is thus a particularly frequently used and popular bioinformatics method for identifying genes or proteins in the genome.

Explain the BLAST algorithm (hint: it is sufficient to describe how the algorithm can become so fast). Also describe its usefulness for biology. If both are still unclear, simply refer to the chapter again.

Task 1.5
Develop a simple program that examines a sequence for possible sequence similarities in a database (hint: enumerate what parts this program would consist of).

Task 1.6
Which of the following statements about BLAST is correct (multiple answers possible)?

A. BLAST = *Basic Local Alignment Search Tool.*
B. BLAST = *Basic Low Alignment Search Tool.*
C. BLAST is an algorithm for finding locally similar sequence segments in a database.
D. BLAST uses a heuristic search and here the two-hit *method* (2-hit method).

Task 1.7
Example: The sequencing of a diseased person has revealed the following protein sequence:
>unknownsequence 1.7

```
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFIKVRQYDQIL
IEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
```

Which BLAST algorithm would you choose for your patient sequence?

A. blastn.
B. blastp.
C. blastx or tblastx.
D. tblastn.

Task 1.8
You now want to know exactly which virus the person has contracted. Perform a BLAST search yourself using the protein sequence (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Which of the following statements is correct (multiple answers possible)?

A. The sequence is almost certainly the pol protein and protease of the HIV-1 virus.
B. The unknown sequence shows low similarity to the pol protein and protease of the HIV-1 virus.
C. When searching for a sequence that is as similar/identical as possible, a match should always have as large an *E-value* as possible and a low identity.
D. The *E-Value* (expected value) shows how likely it is that the hit will be found again in the database with a similar or better score.