A Guide to Molecular Sequence Analysis

Introduction

This guide will introduce the reader to molecular sequence analysis. In the context of this guide, sequence analysis is the process of trying to find out something about a nucleotide or amino acid sequence, employing in silico biology techniques. You may have sequenced a gene yourself, and wish to learn what the long string of letters representing bases, actually code for. You may want to confirm that you have indeed cloned a gene successfully, or you might want to learn about a sequence of DNA that you know absolutely nothing about. You may want to know if a worm has a similar protein to a human one. These, and many other situations, require that you employ sequence analysis.

How It’s Done

Vast databases of genetic information have been made publicly available. Some information is not made public, and you may have access to such databases at your place of work. Currently, many of the international scientific journals require that a sequence be submitted to a publicly available database, before the discovery of a new sequence can be published. Large numbers of sequences have been checked and published, often annotated and cross-referenced. Each record (entry) is curated and maintained in one of the many different databases accessible over the Internet.

Software is freely available that will compare your unknown sequence to all of the sequences in a database. Sequences which are similar to yours are reported. You may also find some utilities useful, in particular those which predict which sequences are coding, or those that present you with a graphical, three dimensional image of a macromolecule.

What’s In This Guide?
**In silico** biology is a thorough, expanding and complex science. This guide provides an interactive working introduction, for scientists with no working knowledge of molecular sequence analysis.

You will learn the essentials of molecular sequence analysis by performing your own searches of provided "unknown" sequences. Each database has its strengths and weaknesses, and you will learn to choose the most appropriate database for your desired search. You will learn a little about the many options available when performing searches.

A detailed explanation of such options, and of the complex statistics applied, will not be covered. Advanced users who wish to learn about these aspects, can refer to the scientific literature ([References](#)). There are also many excellent online guides for advanced users, which can be found through a simple Internet search.

If any of the links fail, please send an E-mail to the webmaster immediately. Comments and suggestions about this guide are also welcome.

**Access to the Guide**

Many links in this guide will take you to another page at this site (e.g. to the [glossary](#)). To return to the referring page, use the back button on your web browser.

---

**A Guide to Molecular Sequence Analysis**

**Contents Page**

**Where to Start**

**Newcomers:** Start at chapter one and read the whole guide. You are advised not to omit sections.

**Some Experience:** You may wish to start at a specific section. If you find that you don't understand a chapter, go back and read the previous one. Each chapter assumes that you have read the previous one!

The interactive exercises require increasing application of your learned knowledge as you progress through the guide. If you have skipped a section, and don't understand the instructions in an exercise, go back and read the previous chapter(s).

An overview at the top of each page links to the major sub-headings on the same page. This may help you to navigate long pages on a small screen.

**The Guide**

The [site map link](#) can be used to locate a specific section in each chapter. Each main heading is listed.
1. Molecular Databases with [Site Map]
   An introduction to the main molecular databases, what they store, how to access them and how to understand the output. Identifying a databases strong and weak points is also discussed.

2. Nucleotide Database Searching with [Site Map]
   An introduction to nucleotide (DNA, cDNA) sequence alignment and database searching. There are numerous applications which require genomic studies; a wealth of biological information is found in the non-coding regions of DNA.

3. Protein Database Searching with [Site Map]
   An introduction to protein sequence alignment and database searching. This is the preferred method for many applications, including studies of molecular evolution; protein sequence comparison is 2-5 times more sensitive than for DNA.

4. Molecular Sequence Alignment with [Site Map]
   How to identify differences in two similar sequences by aligning them. Users with access to GCG can also learn how to use seqed and bestfit.

5. Further Sequence Analysis
   Brief introductions to other tools of bioinformatics for readers who want some pointers of where to go from here.

The Appendix

1. Useful Links (Full List)
2. Useful Links (Concise List)
3. FASTA Format Explained
4. The Genetic & Amino Acid Codes
5. References
6. Glossary of Terms and Abbreviations

Comments, corrections and reports of failed links should be reported to the webmaster.

Molecular Databases

General Introduction

Welcome to the chapter on molecular database searching. This is the most important chapter of The Guide to Molecular Sequence Analysis. Fundamental concepts of in silico biology will be explained, and you will learn how to use and understand the databases that store the wealth of information that is so useful to the molecular biologist.

If you don't understand a word or abbreviation at any time, check the glossary. A link to the glossary is provided at the bottom of every page.
Feel free to recommend new words and abbreviations, or indeed corrections if you disagree with my explanation!

In this chapter, I refer mostly to DNA database searching, but the principles are the same for proteins. Protein databases are discussed later in this guide.

Let's imagine that we have made a cDNA library, and that we have sequenced some of the cDNA clones. We don't know anything about these sequences, and looking at the string of letters representing nucleotides (A, C, G, T) doesn't give us many clues. We certainly cannot tell what the in vivo function is.

Now let's consider an analogy: a word is another sequence of letters that can either mean a lot (if you know which word the letters spell), or nothing at all (if you don't know the word). If I were to present you with an English word, that you had never seen or heard before, you would be able to find out what it meant by looking it up in a dictionary. This is what we're going to do with our molecular sequences. We are going to look up our unknown sequence, and try to find out more about it. We will search for the sequence in the vast molecular databases available on the Internet.

**Introduction to Molecular Databases**

Returning to the analogy of an unknown word, suppose you didn't know anything about the word "nucleus". You could look it up in almost any English dictionary and find an explanation. If you wanted to see a picture of a nucleus and learn even more about it, you would look in a biology text book or an encyclopedia.

If the word was "contig", you might find you have to try a little harder. A good dictionary will list "contiguous" (next in order, neighbouring), but not contig. Since the words are similar, you might postulate that "contig" is derived from "contiguous" (or vice versa). Such a relationship between two sequences might be described as homology. If you didn't find the word in a dictionary, you could search other, more specialised sources such as text books and scientific journals. You would be more likely to find an answer there, but it might take longer to find.

Applying the word analogy to our unknown genetic sequence, let's consider how we should go about identifying it. We should start by searching in a general database (analogous to a dictionary). If our search was unsuccessful, or only partly successful i.e. if only sequences with some similarity are found, then we would use any information we could, and search a more specialised database. If, for example, we found that most of the similar sequences are related to the immunoglobulin supergene family, it would be wise to look for our sequence in a database that specialises in immune genes.

Only after exhaustive failed searches can you consider whether your sequence is a new sequence that has not previously been described. Even if you have discovered a new sequence, you might be able to find related sequences from other organisms. The more closely related the organism, the more similarity you should expect. You might be able to learn something about the possible function of the gene.

When we are satisfied that we have identified the sequence, we can follow links to (or search) other databases which are analogous to encyclopedias or specialised text books, providing a wealth of related information. Frequently, this information is cross referenced with articles published in the international scientific journals.

**Database Searching Options**
Let's get an overview of how to search the molecular databases to identify a query sequence, before you do some searching of your own. Don't worry if this seems complicated at the moment, it will make more sense as you go through the interactive exercises.

Unfortunately, searching for nucleotide or protein sequences is not quite as simple as searching for words in a dictionary. Just as the word "colour" is spelt "color" in American English, so sequences may vary. An equivalent sequence may be slightly different in one organism when compared with another, and mutations can give rise to differences between sequences from identical regions in different cells. You need to be aware of this when searching the molecular databases.

sequence Alignment

You can search for sequences that are just a few bases in length, or many kilobases long. As you may not be looking for a whole "word", the databases are searched by attempting to align your unknown sequence against those on record. Imagine sliding one past the other, until there is a match.

Gapped Sequences

Your sequence may have a region missing (deletion) or a new region may have been inserted into it as a result of a mutation. When searching the databases, you will sometimes need to identify sequences where the alignment allows for a biologically significant gap (inserted or deleted region) in the sequence.

This is also important when aligning cDNA, essentially a concatemer of exons, against genomic DNA. The human genome comprises only 5% exons, interspersed with introns. Computer programs have been developed, which apply advanced statistics to the database search, allowing even gapped sequences to be aligned.

Alignment Programs

Different computer programs are available to perform database searches, with varying levels of search sensitivity depending on the statistical algorithm that they employ. As you will learn, each program has it's advantages, and it's disadvantages. Basically, the more sensitive programs are slower.

Substitution Matrices

With a little thought, you will realise that the number of nucleotides in the database that must be compared to your sequence, is quite vast. Advanced statistical matrices (References) have been developed, that allow a query sequence to be aligned with matching sequences in the database, very rapidly. The most significant matches (successful alignments) are reported. The less complex, faster matricies sacrifice a certain degree of match significance i.e. you need a better match for it to be recognised than if you use a slower, more complex matrix. The matrix, together with the choice of program essentially determine the search sensitivity and search speed.

Filters

Filters are sometimes provided, and are usually turned on by default. The filter masks regions of the query sequence (your sequence) that have low compositional complexity, as determined by the SEG or dust programs (References). Masking is achieved by replacing the sequence with a string
of N's (NNNNNN), the IUB code for any DNA base. Refer to the literature for more information. (References).

It is a good idea to mask almost all sequences, if you have the option to do so. Poly-A tails, for example, can give rise to artificially high scores and therefore misleading results. This is due to the large numbers of such sequences distributed throughout the genome, and therefore throughout the database.

Summary

We have learnt that to identify our unknown sequence, we can perform a sequence alignment search to see if our sequence is listed in the public databases. First, we need to choose an appropriate search program. We need to decide which database is most appropriate to search, and which matrix will give a rapid, accurate result. If the program allows, we also need to decide whether or not to allow for gaps in the sequence, and if sequences should be filtered.

Making the right choices will help to ensure that you have the best possible chance of getting a relevant and accurate answer, quickly.

An Overview of Public Molecular Databases

The Main Databases

There are currently three main public databases: (Abbreviations are explained in the glossary)

1. EMBL: at the European Molecular Biology Laboratory, Cambridge, UK.
2. GenBank: at NCBI, a division of NLM at the NIH campus, USA.
3. DDBJ: the DNA Databank of Japan.

These three databases have collaborated since 1982. Each database collects and processes new sequence data and relevant biological information from scientists in their region e.g. EMBL collects from Europe, GenBank from the USA.

These databases automatically update each other with the new sequences collected from each region, every 24 hours. The result is that they contain exactly the same information, except for any sequences that have been added in the last 24 hours. This is an important consideration in your choice of database. If you need accurate and up to date information (e.g. if you plan to publish), you must search an up to date database.

This guide will frequently make use of the EMBL & GenBank databases. When you have completed the exercises in this guide, you may like to search all three databases. You will be able to see for yourself that the results are the same, even if reported in different fashions. You may find that you prefer the presentation of one database. You can safely continue to use your favourite database, knowing that the contents are identical (to within 24 hours).

In 1998, there were more than 1200 million base pairs from over 1.6 million sequences in the EMBL database, and the doubling size was estimated at only around one year! To organise this huge amount of information, the database has been split into numerous divisions (17 in 1998). Each entry (sequence and relevant information) is allocated to exactly one division. The division is indicated by a three letter code, specified when you receive information about a sequence. The codes are given below:

Remember to refer to the glossary if you don't recognise abbreviations!
Coding of Database Entries

Each entry in a database must have a unique identifier that is a string of letters and/or numbers that only that record has. The identifier can be quoted in scientific literature, as it will never change. As the unique identifier must always remain the same, another code is used to indicate the different versions of a record: new information might be added, or a sequence might be corrected, for example. You should therefore always take care to quote both the unique identifier and the version number, when referring to records in a molecular database. The entry identification codes for EMBL and GenBank are described below:

- **EMBL**
  - **Identifier (ID)**
    This number is the unique identifier. It is allocated when a sequence submitted to the EMBL database is accepted for publishing. It will never change. It will be quoted in an EMBL report and may be quoted in the description line of a FASTA report.

    Example:
    - In an EMBL Report - HS498971

  - **Accession Number (AC)**
    This is the unique identifier allocated by GenBank. See the description for GenBank Accession Number below.

    Example:
    - In an EMBL Report - AC U49897

- **Nucleic Acid Identifier (NI)**
  This number is assigned to each version of an entry. While the identifier (EMBL) and accession number (GenBank) never changes, a new NI number will be allocated each time the sequence is modified, however minor the change might be.
Example:
- In an EMBL Report - NI g2462721

● GenBank
Accession Number (AC)
The accession number is allocated when the record is first entered into the database, and will never be changed. It consists of one letter followed by 5 digits (X12345), or (more recently) two letters followed by six digits (XY123456). This number is also reported in EMBL reports.

Examples:
- In a GenBank Report - ACCESSION: U49897
- In a FASTA Report - gb|U49897

GI Number (GI, NID)
This number is referred to as the NI number in EMBL --they are identical. The GI number tracks versions of an entry, and was until recently, only quoted on the NID line in a GenBank report. Database collaboration efforts have prompted a change: the GI number is now quoted in a new line called VERSION. Although now redundant, the NID line will remain for quite some time.

Examples:
- In a GenBank Report - NID: g2462721
  [NID: Now redundant.]
- In a GenBank Report - VERSION: U49897.1 GI: 2462721
- In a FASTA Report - gi|2462721

Accession.version
This is a new field, and the one to pay most attention to. The first number is the never changing accession number, followed by a period and a version number. The version number starts at one, and increases by one each time the sequence changes. The second number is the GI number (see example under GI Number, above).

Example:
- In a GenBank Report - VERSION: U49897.1 GI: 2462721

For the purpose of this guide, you will find it useful to remember that the EMBL ID and GenBank Accession numbers are unique. It is possible to search these databases, and others (e.g. protein databases) quoting just these unique identifiers.

Other information is stored along with the sequence. Each piece of information is written on its own line, with a code defining the line. For example, DE, description; OS, organism species; AC, accession number etcetera. Most are self explanatory from the content. Sometimes, a user friendly graphical report is made available, making it even easier to read the results of a search. Relevant biological information is usually described in the feature table (FT).

Other Databases
Research trends and other factors have influenced the content of molecular databases. Some databases have specialised, and you will find a couple of these listed in the list of links provided. If you are interested in immunogenetics, for example, you would find that the IMGT (International Immunogenetics Database) is more comprehensively annotated than the general databases such as EMBL and GenBank.

There are so many specialised databases, that it is not reasonable to address them here. When you have completed this guide, you will be able to
apply your knowledge to almost all molecular databases on the Internet. When using other databases, take care to check how regularly updated the information is, and note whether or not the database is redundant.

**Database Redundancy**

A redundant database is a database where more than one copy of each sequence may be found. No effort has been made to reduce the inevitable redundancy (duplication of sequences). Such a database is more comprehensive and more likely to contain recently discovered sequences. The penalty paid is that biologically significant results are more likely to be hidden amongst the large numbers of irrelevant reported matches.

Non-redundant databases are less complete, but the reduction of redundancy should make finding a biologically significant sequence easier and more successful. Non-redundant databases are frequently referred to as *nr*.

The term is ambiguous with reference to molecular genetics, and as such, the degree of non-redundancy varies according to the database manager's interpretation. One can argue whether or not two alleles of a locus defines the limit of redundancy, or whether the same locus in different, closely related organisms constitutes redundancy. Also, whether two perfectly aligned sequences should be the exact same length to be regarded as identical.

The concept of redundancy is important, if you are to have a high degree of success when searching the molecular databases. You want to search a database that will return possible match sequences of biological significance, but not too many of them. If too many matches are reported, the most biologically significant sequence might "drown" in the sea of suggestions.

---

**Nucleotide Database Searching**

**Sequence Alignment**

The first decision you make when searching molecular databases is which program to use. There are three main different programs available. The choice of program to use is highly debated (similar to the Mac vs PC debate). It is generally recommended that you use a BLAST program first; BLAST is the fastest option and yields good results. If the results are not satisfactory for your search, use the FASTA program. This is a good, general rule of thumb. You will learn with experience, which program is best to use with a certain type of sequence.

BLAST is the fastest sequence alignment program, but compromises some degree of sensitivity in favour of speed. FASTA is slower, but more sensitive. You may like to try, BLITZ, which also provides a sensitive search, but is slow to run (i.e. demands a lot of computer processing power). You may be able to run BLAST locally, if it is provided to you by EMBnet for example. Alternatively, you can use any of the numerous bioinformatics servers which provide a free public service. There are also different flavours of sequence alignment programs, such as BEAUTY, a modified BLAST (a link is provided in the list of links).

Usually, adjusting the search options for BLAST or FASTA is sufficient to
locate your sequence, assuming it is listed in the database being searched! BLAST and FASTA are arguably the most commonly used sequence alignment programs, and this guide will only explain these.

When you have completed this chapter, you will know enough about molecular databases and their corresponding search programs, to allow you to use any other molecular database or search program that you may find. The principles are generally the same, even if the methods are different.

**Sequence Alignment with BLAST**

*(Basic Local Alignment Search Tool)*

BLAST is the algorithm used by a family of five programs that will align your query sequence against sequences in a molecular database. Statistical methods are applied to judge the significance of matches. Reported alignments (i.e. sequences in the database that could be identical to your query sequence) are reported in order of significance, as estimated by the applied statistics. The BLAST algorithm has been optimised for sequence alignment, but not for motif-searching.

Listed below are the definitions for each BLAST flavour, as described by NCBI (with permission):

**BLASTN**  
Compares a nucleotide query sequence against a nucleotide sequence database.

**BLASTP**  
Compares an amino acid query sequence against a protein sequence database.

**BLASTX**  
Compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

**TBLASTN**  
Compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).

**TBLASTX**  
Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The BLAST programs are the fastest currently available. The speed was originally achieved (version 1.4) in part by forbidding gaps in the sequence. As mentioned before, gaps affect the quality of your results. Most servers now offer BLAST version 2.0 (BLAST2.0), which can perform gapped alignments. The search is accelerated by the heuristic nature of the statistical algorithm used: BLAST performs fast, local rather than slower, global alignments. BLAST does not try to match the whole sequence.

**BLAST Exercises 1**

Now that you have quite a solid background for sequence analysis, it's time for you to run a search. This first search will give you a feel of what it's all about. Afterwards, we will look more closely at some of the details of molecular sequence alignment.

You may find it helpful to write down (or print) the instructions. You should complete this task before proceeding—it won't take long, and will
greatly enhance your understanding.

If you get stuck, have a look at a [screenshot](50.6 Kb) of the EXPASY BLAST page.

**Basic BLAST**

**Usage:** Choose the the suitable BLAST program and database for your query sequence. Paste your sequence in one of the supported formats into the sequence field below and press the "Run BLAST" button. Don't forget your e-mail address, so that we can send you the results in case of traffic jam...

Make sure that the format button (next to the sequence field) shows the correct format.

See also our [BLAST database description](#) and the announcement on the [TIGR web server](#).

**Please select the program:**

- blastn

**Please select the database:**

- embl without ESTs (DNA)

**Gapped alignment on off**

**BLAST filter on off**

**Graphic output on off**

**Paste your sequence here:**

(or ID or accession number)

(required for tblastn) programs ->

**Run BLAST**

**Clear form**

---

**Scenario:** You have been given a human gene to sequence and identify, but no clues as to what it is. The provider wants an unbiased opinion. To identify the sequence, you should:

1. Copy the human query sequence, given below:

```
AAAAGAAAAAGTTGAGAAATGAGATGATAAAAGGGGCCATTTTAGGTTTAGGTAATTAAGGGGCAATTTTATGTTAATAAGGTGTAATCTTTACCTGTATACAGTGCAGAGCCTTCTCAGAAGCACAGAATATTTTTATATTTCCTTTATGTGAATTTTTAAGCTGCAAATCTGATGGCCTTAATTTCTTTTTGACACTGAAAGTTTTGTAAAAGAAATCATGTCCATACACTTTGTTGC
```

---
2. Go to the EXPASY (EMBnet) BLAST Server WWW page
   If for any reason, you cannot access the EXPASY BLAST server, you can use any other BLAST server. I will refer specifically to the EXPASY server options and page layout.

3. Select the program: BLASTN
   This is the BLAST program that will compare a nucleotide query sequence against a nucleotide database.

4. Select the database: EMBL without ESTs (DNA)
   This is the main EMBL nucleotide database.

5. Ignore the matrix option.
   It is not used by BLASTN.

6. Select sequence input format: Plain Text
   You will be submitting the nucleotide sequence in plain text.

7. Select the following:
   - Gapped Alignment: ON;
   - BLAST filter: ON;
   - Graphic Output: ON.
   These are all ON by default.

8. Paste the query sequence into the specified area.

9. Hit the button: Run BLAST

10. Wait as your query is processed by the server.

11. Examine the output.

BLAST: Reading the Output

The report might look daunting, but the most important aspects are quite easy to grasp. At the top of the page is a description of the program used (e.g. blastn), the version of software and related information. Next, you will see the query sequence that you entered and BLAST has tried to align. You will see a string of N's if any region has been masked by the optional filter. The next few lines provide information about the database you searched, including when it was last updated.

The last part of the introduction is a series of 50 periods (.) between the words searching and done. Asterisks (*) on this line indicate that the program had difficulty making progress through the database, and fewer than 50 periods indicate that the program was unable to search the whole database (it may have crashed). This should be taken into consideration; you may wish to run the query again.

Below the "Searching...Done" line, you will see a drawing. As you have probably guessed, the top red line represents your unknown, query sequence. The line has a nucleotide scale drawn on it. Below the scale are numerous lines of different lengths and colours, each representing one of the matching sequences from the database. You should examine this closely yourself, but for now, just note that there is only one red line at the top (below the query line) that is as long as the query sequence. All of the lines below it are much shorter. This illustrates in a graphical way that there is just one sequence that shows a high degree of homology with your query sequence. A screenshot of the image (30.8 Kb) is available, in case you cannot access the server.
Further down the report, you come to a list of one line descriptions of database sequences that produced a significant alignment. The first one in the list (represented as the top long red line in the graphic) reads as follows:

```
emb|L37747|HSLAM11 [Homo sapiens]Homo sapiens lamin B1 gene, ex... 416 e-114
```

The *E* value is important. This the last number on the line given above (e-114). The *E* value is given in scientific notation. In the example, e-114 can be read in full as 1 times 10 to the power of minus 114. Or in other words, very close to zero indeed! This value is the number of times you would expect to see such a match (or better) merely by chance. The closer the value is to zero, the less likely the event is. To reiterate, the smaller the *E* value, the more significant the match is.

We know that the sequence we tested is human, so there are only two alternatives from the list: the first and second. The others are not only poor matches with *E* values indicating that these matches are probably the result of chance, but most of them are also from other organisms.

A word of warning! Lamin B has a special (uncommon) nucleotide sequence. When you search for your own sequences, you are likely to get more than one optimal sequence reported. The most statistically significant match (lowest *E* value) is not necessarily the sequence you are looking for! Look out for short repetitive sequences, and motif families that may not be accounted for by the statistical alignment algorithms (e.g. BLAST). Use the reported significance as a guideline only. Your common sense combined with your knowledge of molecular biology and the sequence source should always be applied when searching the databases.

Yet further down the report, we can examine the actual alignments that were found. These are listed in the same order as the short descriptions. A short example of an alignment is given below:

```
Query:   1 ggccccaccacgccgctcag 20
          ||||| ||||| |||||           |
Sbjct:  701 ggccccaccacgccgctgag 720
```

We can see that there is almost 100% alignment between the query and subject (database) sequence. A bar (|) joining two nucleotides indicates homology. The 18th query nucleotide is a C which is different to the database sequence.
This is shown by the absence of a bar (|) at that position. If you look at the alignments of the other reported sequences, you will notice that there are even more gaps.

Non-matching of nucleotide pairs may represent the absence of homology between the sequences. It can, however, also be due to a point mutation in the query sequence DNA (which might be interesting), a sequencing error (gel compression for example; resolve this by sequencing in both directions and aligning the sequences), or even an incorrectly reported entry in the database. There are several other situations where mismatching arises.

The results of this search make it pretty clear that the first sequence is identical to our sequence. Returning to the one line description list, you can click on the EMBL identifier (HSLAM11) and view the EMBL report. This includes relevant information about the sequence, including the protein sequence that it encodes.

Well done! You have now successfully searched a molecular database!

**Understanding BLAST**

Since you will probably use BLAST more than any other sequence analysis software, let's learn a little bit more about what we just did, how it was achieved, and what else we can do with BLAST.

If you return to the EXPASY [BLAST page](#) and hit the button labelled "Advanced BLAST", you will be presented with many more options. The optimal options for most searches are set by default, but you can easily change them for your BLAST query. Before we learn about a few of the most important options, we need to understand the basic concept behind how the search is done.

**Scoring: Finding the Best Alignments**

The BLAST programs attempt to align short regions of your sequence with regions of sequences in the database. The initial scanning phase identifies matching [query:database] fragments. A match is determined by the sum alignment score for a region (defined as a "word") of the query sequence.

The alignment for each base in the word is scored: if a nucleotide in the query word exactly matches a nucleotide at the same position in the database word (e.g. A with A), then a positive score is awarded. If the match is good but not perfect (e.g. W in the query with A or T in the database) then a lower score is awarded. If two nucleotides do not match, a negative score is awarded. The sum score is used to determine the degree of similarity.

Sequences with a high score are referred to as high-scoring segment pairs (HSPs). The program tries to extend the best HSP (those with the highest score; the best matches) by extending the alignment in both directions. The alignment extension is continued until the sequence ends, or the alignment becomes non-biologically significant. Substitution matrices are used during both scanning and extension. The reported sequences are those with the overall highest scores (maximal-scoring segment pair, MSP).

**The WORDLENGTH Option**

The length of the initial word identified is specified by the value W (WORDLENGTH). BLAST only attempts to extend aligned fragments that are a perfect match for W continuous nucleotides. The default is 11 for blastn; by default, blastn will scan the database until it finds words that are 11 letters long that exactly match an 11 letter word in the query. These will be extended. A WORDLENGTH of eleven is sufficient to exclude even moderately diverged homologs, and therefore also excludes almost all chance alignments.
The Filter Option

BLAST version 2.0 enables the application of a filter. The filter masks regions of the query sequence that have low compositional complexity (e.g. Alu sequences), as determined by computer programs (SEG or XNU); (References). Masking is achieved by replacing the sequence with a string of N's (NNNNNN). N is the IUB code for any DNA base. Only the query sequence is masked. The sequences in the database will not be masked.

Filtering is a good idea for almost all sequences, and is turned on by default. Poly-A tails and proline rich sequences, for example, can return artificially high scores and therefore misleading results. This is due to the large numbers of such sequences, dispersed throughout the genome, and therefore also throughout the database.

The Matrix Option

As discussed above, significance of a match is determined by the returned score. The score reflects the probability that this match would not have been found by chance. The method for calculating and optimising the score is the essential difference between the programs and matrices.

Statistical matrices are used both to identify sequences in a database, and to predict the biological significance of the match. There are several to choose from, but you should usually accept the matrix recommended by the program you are using. To understand when to use different matrices, you should understand the matrices and how they work. If you wish, you can review the literature (References), or refer to Keith Robisons homepage.

Briefly, there are two main types of substitution matrices that are frequently used by programs such as the BLAST family. Substitution matrices function to give a score to the alignment of each pair of residues. This isn't as simple as it may seem: remember that IUB ambiguity codes can be included in the sequence, and biological significance is the goal.

For example: a mutation may result in lysine being translated into a protein, instead of arginine. This may be a conservative mutation, as the function of the protein is unlikely to change.

Common Substitution Matrix Families

PAM (Percent Accepted Mutation)
PAM matrices are most sensitive for alignments of sequences with evolutionary related homologs. The greater the number in the matrix name (e.g. PAM40, PAM120), the greater the expected evolutionary (mutational) distance. You should choose the appropriate matrix for an optimal search. If the mutational distance is unknown, you should run at least three searches using PAM40, PAM120 and PAM250 matrices. You may choose to use PAM to identify conserved sequences or features therein, or to establish the evolution of a sequence.

BLOSSUM (Blocks Substitution Matrix)
BLOSSUM matrices are most sensitive for local alignment of related sequences. The BLOSSUM matrices are therefore ideal when trying to identify an unknown nucleotide sequence. BLOSSUM62 is a good general matrix, set by default for protein BLAST searches.

You should use BLOSSUM62 (set by default) for protein BLAST searches i.e. BLASTP, BLASTX, TBLASTN and TBLASTX. BLOSSUM62 is optimised for general BLAST searches, and is suitable for most situations; it will recognise some amino acid substitutions as conservative (e.g. Arg to Lys). If you are searching for evolutionary related proteins, you
should use PAM120 for generalised similarity searches. Take care! You cannot compare the alignment scores (see later) from one matrix directly against the alignment scores from another matrix!

You can choose an alternative scoring matrix for BLASTP, BLASTX, TBLASTN or TBLASTX. You can choose between PAM30, PAM70, BLOSUM80, BLOSUM45, or the default BLOSUM62. You cannot choose a matrix for BLASTN searches (instead, specify M and N, discussed below).

The EXPECT Option

You may, for example wish to set an expected score threshold (EXPECT) for the search, set to 10 by default. This means that ten matches are expected to be found by chance. If the statistical significance of a match is greater than the expected score threshold, it is not reported. Only if the statistical significance is less than this level, will the match be reported. In other words, a lower EXPECT threshold applies a more stringent search. This leads to fewer chance alignments being reported. You can enter fractional values if you wish; values are often suggested in a menu.

The Score Value Options

At the top of this page, we learnt that a [query:database] nucleotide pair was rewarded with a score depending on whether the nucleotides at that position were identical or not. The score awarded can be set by the user.

M Parameter
The score awarded when a pair of aligned residues match. Must be a positive integer.

N Parameter
The score awarded when a pair of aligned residues do not match. Must be a negative integer.

The ratio of M:N determines the degree of divergence (evolution) that is accepted. The default value for M is 5 and for N is 4. The ratio of 1.25 equates to around 47 nucleic acid point accepted mutations (PAMs) per 100 residues. PAMs are used as a predictor of the degrees of evolution from an ancestor (in molecular terms). If you adjust the M and N values to give a higher ratio, more nucleic acid PAMs will be accepted by the algorithm, resulting in a more divergent search.

Fetching Sequences

Finally, you may have noticed that you don't need to enter a nucleotide sequence when performing a BLAST alignment. You can enter the ID number (EMBL entries) or accession number (GenBank) instead! This method allows you to fetch a sequence from the database, without performing an alignment. There may be several reasons you want to obtain the latest accurate version of a sequence. You may be optimising a PCR, and need to confirm the quality of your PCR reaction by sequencing your PCR product, then comparing your sequence against the published sequence in a database, for example.

There are yet more options, but they require a more complete understanding of the search algorithm to be applied successfully. I suggest that you read the manual pages at a BLAST server site for more information.

BLAST Exercises 2

Let's get some practice at BLAST searching, with a couple of examples. If you get stuck, go back and glance through the last page or two again. The
instructions are all there.

You should use the EXPASY **BLAST Server** for these searches. Write down your answers to the proposed questions, and check them against the answers given along the way.

**Exercise 1**

Copy the sequence given below, and use it to run a nucleotide BLAST search. Use the default criteria except where you must change the options. Hint: You may need to select a program and database! Can you identify the sequence?

GTCCGGCCTGGGCGACAGAGCAAGACTCGTCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

**Exercise 2**

The sequence is taken from a GenBank entry whose accession number is S56967. Using the BLAST server, locate this sequence.

Seeing the actual sequence, you shouldn't be surprised that you were unable to identify it in your first search... it's an **Alu sequence**. No wonder there were so many good matches in the first search! If I hadn't told you that it was part of an Alu sequence, you might not have guessed, but you would have been able to make the discovery yourself. Read on...

**Exercise 3**

Copy the above sequence again, but this time use the blastx program. This program translates the nucleotide sequence in all six reading frames (both strands) and compares the result to proteins in the SWISS-PROT protein database; a major European protein database.

Paste the sequence into the BLAST query form, select **blastx** from the program menu and **SWISS-PROT (PROT)** from the database menu. Run the search and examine the results. The layout is slightly different, but principally the same. From the results (only 2 when this guide was written), you can clearly see that the sequence is an Alu sequence. You would therefore need to sequence a different region of the gene, or a longer stretch, in order to identify it.

**Exercise 4**

Copy the sequence below, and run a BLAST search. Examine the results.

GAATTCTAATCTCCCTCTCAACCCCTACAGTCACCCCATTTGGATATATAAGATGTTGCTACTGTCTATGTATCCCTCAAGTAGTGTCAGGAATTAGTCATTTAAATAGTCTGCAAGCCAGGAGTGGTGGCTCATGTCT

You should be able to identify the sequence as coding for human beta haemoglobin. When this guide was written, there were two alignments that were not only very good matches (100% and 99% homology), but they matched for the whole length of the query sequence. The remaining reported sequences were very short indeed. It is quite clear cut that the gene is a beta haemoglobin gene, but the second sequence has a C instead of a T. You would be wise to double check your sequencing before accepting the identity.

**Exercise 5**

The sequence in the last alignment is clearly quite unique. What would happen if you didn't know any more than the first 15 bases? Try it! Here they are:
Nothing found? That seems rather odd, because we know that there is a corresponding sequence in the database --we found it in the previous search! The point of this exercise was to demonstrate that although the database returns a "No Hits" message, this may not be because there isn't a sequence to be found. It may be because the search criteria need to be optimised for the search. Some of these options were discussed earlier.

Copy the 15 nucleotide sequence again, and return to the BLAST page. This time, click on the "Advanced Blast" button. Use the default parameters, except you should select the EMBL database, and nr from the list of EMBL divisions. Now adjust the E value to 100. You may remember that doing so will reduce the specificity of the search by accepting more chance alignments. You should also turn off the Xblast-repsim filter. Failure to do so will mask almost the entire sequence with N's! Now hit the run button.

When this guide was written, only two hits were returned, both of which were the same two most significant alignments from the previous search, identifying this sequence as being most likely to be a beta haemoglobin gene.

Well done!

Finally, you may like to try and run the last search once more, but with an even larger E value, say 10 000. You will get lots of "significant" alignments reported. The two reports we have seen so far may well be at the top of the list, but we cannot exclude all of the other alternatives, with the information given. Too many chance alignments have been accepted.

**BLAST or FASTA?**

If the last sequence were further shortened by three bases, to read GAATTCTAATCT, you would not get any (useful) reports from the Basic BLAST, or by increasing the E value in the Advanced BLAST. You may like to try adjusting other options in the Advanced BLAST to see if they make a difference. Alternatively, we could follow the advice given earlier: when you have trouble finding a sequence with BLAST, try FASTA.

**FASTA Exercises 1**

In the last exercise, we found that BLAST was unable to suggest a database entry for GAATTCTAATCT, even though we knew that the parent sequence did have an entry in the database. The reason it could not be found (at least not without some advanced option tweaking) is largely due to the sacrifice BLAST makes to search sensitivity, in favour of speed.

FASTA is another fast sequence alignment program, that is more sensitive than BLAST. You would be right to assume that the increased sensitivity incurs a speed penalty. You can be left waiting for the query to report results for some time, depending on server demand and the search you are running. Although a search can take hours, it is usually just a question of minutes. The prolonged processing time is a good reason in itself, for choosing to use BLAST first.

**Exercise 1**

Now that you know a bit about sequence aligning, you can easily run a FASTA search for the given sequence. The instructions are given below:

1. Copy the 12 nucleotide sequence: **GAATTCTAATCT**
2. Go to the [GeneStream FASTA home page](http://www.genes.com)
3. Paste the sequence into the specified window and run the search using
the default parameters.
4. Be patient! If your query takes time to process, please don't close the
window and give up --unless you really must. The server will still be
processing your query.
5. Examine the output. Could you identify the sequence if you didn't know
what it was?

Even with the short sequence provided, there were many alignments reported.
Did you notice how the reported alignments were presented differently to the
BLAST results? If not, you might like to go back and look again. Notice that
BLAST only reports the aligned bases, whereas FASTA also presents the
context of the alignment i.e. the flanking sequences on either side.

You will probably have realised that the FASTA reported sequences were
almost all clone or vector sequences. When this guide was written, there
were only two reported sequences that were not clone or vector based: human
beta haemoglobin and gibbon gamma globin. As we can be certain that our
human subject won't have donated gibbon DNA, then we can successfully
identify the sequence, even from just twelve nucleotides! Impressive, isn't
it?!

Understanding FASTA

The FASTA algorithm and family of programs are similar to BLAST in that they
both align a query sequence against all of the sequences in a database and
return the most significant matches. Whereas BLAST relies on the sum match
probability for each local alignment for the sequence, FASTA scores only
exact matches. FASTA allows gapped searches to be made. Like BLAST, FASTA is
heuristic, sacrificing some speed for sensitivity.

FASTA comes in several flavours, and you should choose the most appropriate
program when searching.

```
fasta3
A DNA query sequence is aligned against a DNA sequence database. A
protein query sequence will be aligned against a protein database.
```

```
tfasta3
Align a protein query sequence against a DNA sequence database,
translating the DNA sequences 'on-the-fly'.
```

```
fastx3
Align a DNA query sequence against a protein sequence database,
comparing the translated DNA sequence in three frames.
```

```
tfastx3
Align a protein sequence to a DNA sequence database. Align with forward
and reverse frameshift mutations.
```

```
ssearch
Align a protein or DNA sequence to a sequence database using the
Smith-Waterman algorithm.
```

FASTA Options

Sequences as short as 10 nucleotides in length can be queried using FASTA.
The speed of the alignment is largely determined by the \textit{KTUP} value, which is
used to limit the word length. You may recall from the BLAST pages, that a
"word" is a short region of the query sequence that is compared against the
database. In BLAST, words with the highest alignment score are selected for
the extension phase. In FASTA, the word is not scored, but must be an exact match if it is to be processed further.

**FASTA Matricies**

When FASTA identifies an exact match, it uses a substitution matrix on the word to determine the optimal alignment of the query sequence against the identified database sequence. The matrix is used to score the process of aligning the sequence flanking the exact match. The main difference between the initial scanning and this extension phase, is that the substitution matrix (here only used for extension, unlike BLAST) allows the IUB ambiguity codes to be included and scored. Finally, FASTA attempts to join continuous regions of alignments.

Most FASTA servers offer a choice of the BLOSSUM or PAM family of substitution matrices (as for BLAST). For FASTA, the default and recommended general purpose matrix is BLOSUM50. BLOSUM62 is also a good choice, if you find that the BLOSUM50 is unsuitable. If you are searching for mutationally (evolutionary) distanced sequences, try PAM.

**The Smith-Waterman Algorithm**

You may have the option of using another algorithm that is more rigorous than those we have discussed to now. The Smith-Waterman algorithm is employed by the ssearch3 program distributed with FASTA, and should be used if you need a highly refined search. It is much slower to use than either BLAST or FASTA programs.

You may find the TimeLogic server is fast for Smith-Waterman searches, but you will need to register to use the site. Registration was free at the time of writing, and came with benefits including access to other specialist databases. This guide will not use the TimeLogic server.

You will usually be required to enter an e-mail address for the results to be returned to you when they are ready. The Smith-Waterman algorithm is highly computationally intensive, and therefore very slow, but it is undoubtedly worth using if you find you need to.

**The KTUP Value and Other Options**

As mentioned above, the KTUP value can be used to set the word length. You can choose between integers 1 and 6, assuming your server has sufficient processing power to cope! Some servers may restrict access to a KTUP of 3-6. A query where KTUP=2 is five times faster than if KTUP=1! If KTUP is set to 1, the server must align each nucleotide, whereas if KTUP=2, it can align in pairs. The default is set to 6.

You may recall the significance of gapped sequences. Regions of sequences may have been mutated in or out of your query sequence, compared with the sequence in the database. You can set the penalty score for gaps (regions of no alignment between regions of successful alignment) according to how interested you are in this phenomenon.

The GAPOPEN option is the penalty score awarded for the first residue in a gap: it penalises the existence of a gap. The value must be negative and is set to -16 by default. The GAPTEXT option defines the penalty awarded for each additional nucleotide in the gap, and is defined as -4 by default. Note: For fastx and tfastx, a penalty for a frameshift can also be set.

The STRAND option is normally set to upper, telling the FASTA program to compare the query sequence as it is, to the database sequences. The other option is bottom. When set, FASTA will reverse and complement the query
sequence, and use the result to search the database.

**FASTA: Reading the Output**

The FASTA output is essentially very similar to the BLAST output. You are first presented with a list of the reported sequences: the most significant alignments. Below the list, the actual alignments are presented in context of the database sequence. The number of bases that match exactly are reported as a percentage of identity.

In the list of all reported sequences at the top of the page, the last but one value is the score, and the last number given on each line is the expect value (scoring E Value). The maximum (threshold) E value is 2.0 by default. As with BLAST, the smaller the expect value, the lower the probability is that the reported alignment is a chance finding. Putting it another way, the lower the reported expect value is for a reported sequence, the more likely it is that it is true.

You must exercise caution when reading expect values, and regard them as guidance tools only for identifying the origin of your query sequence. You may recall from the FASTA search you did (or you may wish to return to FASTA Exercise 1), that the true reported sequence, human beta globin, was reported with an expect value of 0.57, following at least twenty one sequences with a lower score.

If the results are reported interactively (i.e. as a web page, as opposed to in an E-mail), clicking on the red ball will take you to the database entry for the complete sequence. Clicking on the sequence code in the list will take you to the alignment, further down in the same page.

**FASTA Exercises 2**

You will use two different servers for these exercises. The GeneStream server is based in France, and the EMBL server is in the UK. The reason for using different servers is to give you a feel for the variety of presentation. You will also notice that each of these two servers allow you to change certain search criteria (such as KPUT, STRAND) but not others. It is important to realise that these are essentially the same programs, it is just the web presentation that is different.

You should have plenty of spare time when you do these exercises. Each search can take from two minutes onwards. Please do not close a window (kill a search) once you have initiated one. Doing so will tie up the server, slowing it down for other users.

**Exercise 1**

Now that you know more about FASTA, let's do a simple search, and identify the sequence. Follow the instructions given below:

1. Copy the sequence:
   
   `CCAGATCCTGGACAGAGGACAATGGCTTCCATGCAATTGGGCAGATGTGTGAGGCACCTGTGGTGACC`

2. Go to the GeneStream FASTA server.
3. Paste the sequence into the query sequence window.
4. Select the database: `gbpri` (GenBank Primates).
5. Hit the Perform Search... button.
6. Be patient! Your results will arrive when they are ready.
7. Examine the output.

When this guide was written, one could easily deduce the identity of the sequence. The first reported sequence presents with 100% identity for a
database sequence. The E value is very close to zero (3.9e-15, or 3.9 times ten to the power of minus 15, or 0.0000000000000015) suggesting that it is unlikely that this match is the result of chance. Furthermore, we can see that our query sequence is the same length as reported sequence number one (GenBank AC: S78558). It appears to be the same sequence.

Clicking on the red ball takes us to the full GenBank entry for that sequence. You should scan it briefly, noting in particular the gene identity: exon DNA of BRCA1 gene. A 68 nt genomic mutant. [present in the database because this mutation may be involved in the genetic predisposition to breast cancer].

Okay, so the first sequence appears to be a likely candidate, but you don't know if this is indeed the sequence that you have been given. The second sequence reported (GenBank AC: HSU14680) is the complete human BRCA1 gene. There is a single point mutation, which has reduced the identity to 98.551%. The E value is still favourable. The next sequence is for BRCA1 mRNA and then other variations of BRCA1, with ever decreasing identities and E values.

I'm sure that even the meticulous Mr Sherlock Holmes would conclude that the true identity of the query sequence is in fact that of the first sequence. We read that the sequence with 100% alignment was a region of the BRCA1 gene (the second listed), but with a point mutation. This is exactly what we have seen, and there is only one sequence with 100% ID.

To be absolutely certain, since the differences between the two candidate sequences are of the single base resolution, you should consider sequencing your query sequence again, to confirm it's identity. This is particularly true if the DNA was sequenced using an automated sequencer: a phenomenon of common dye chemistries is that the G signal is very weak, and can be misinterpreted by base calling software. As it is a G that is mutated, it might be wise to sequence the fragment in question, again.

**Exercise 2**

If you have a cDNA sequence, you may wish to align the complementary (genomic) strand. Rather than translating it yourself, you can ask the FASTA server to reverse and complement the query sequence, 'on-the-fly' (i.e. as part of your search). To illustrate the importance of aligning the correct strand, follow the instructions below to align the same sequence as you did in exercise 1, but using the bottom strand.

1. Copy the DNA sequence:
   ```
   CCAGATCTGGACAGGACAATGGCATTCCATGCAATTGGCGAGATGTGTGAGGCACCTGGTTGACC
   ```
2. Go to the EMBL FASTA server.
3. Paste the sequence into the query sequence window.
4. Select the fasta program: fasta3
5. Select the database to search: EHUM (EMBL Human)
6. Select the STRAND: bottom
7. Hit the Run FASTA3 button.
8. Be patient! Your results will arrive when they are complete.
9. Examine the output.

When this guide was written, only two sequences were returned, and both had an identity of less than 70%. Neither look like a good match. You should consider aligning the other strand, if you get results like this in the future. If you had just accepted one of these sequences, you might have thought that the BRCA1 sequence was a Down Syndrome related sequence. That would be a serious mistake!

**Exercise 3**
Finally, let's try the fastx program, which translates the DNA sequence 'on-the-fly' into the three reading frames of the given strand, and compares the amino acid residue sequence with the entries in a protein database. We will use the same sequence as elsewhere on this page. Follow the instructions given below:

1. Copy the DNA sequence:
   CCAGATCCTGGACACAGGACAATGGCTTCCATGCAATTGGGCAGATGTGTGAGGCACCTG
   TGGTGACC
2. Go to the GeneStream FASTA server.
3. Paste the sequence into the query sequence window.
4. Select the fasta program: fastx
5. Select the database to search: SWISS-PROT (Protein database)
6. Hit the Run FASTA3 button.
7. Be patient! Your results will arrive when they are complete.
8. Examine the output.

By examining the output, you will have noticed two things. Firstly, the aligned sequences are in **amino acid one letter codes** and secondly that where aligned pairs have similar properties, but are not identical, a period (.) replaces a colon (:) that indicates homology when drawn between the two bases.

In the specific example, when this guide was written, there were only two likely candidates returned, presenting with good identities and very low E values. The third sequence only showed 50% identity. Not surprisingly, the first two were the protein sequences corresponding to the two nucleotide sequences we identified and discussed earlier. Using fastx, not only have you gained even more strong evidence that the correct sequence is a BRCA1 mutant, but also shown that the sequence is coding, and you have found the protein it has coded for.

You started with a 68 nucleotide long DNA sequence, of which you knew nothing whatsoever. You have now identified the sequence, found an entry in both GenBank and EMBL (identical, remember!) with citations to the scientific literature and other information, and you have the protein sequence. In fact, you can even go further, and see a three-dimensional image of the protein, that you can spin on your screen. This, and more, is covered in **Protein Database Searching** chapter of this guide.

Congratulations! You have now completed the most comprehensive chapter: nucleotide sequence analysis. The next chapter will be about nucleotide sequence comparison.

---

**Protein Database Searching**

**Introduction**

STOP! An important note from the author. This chapter assumes previous knowledge! If you have not read: Molecular Databases, you should do so before reading any further.

Protein database searching is the most important method to master. It is between two and five times more sensitive than DNA database searching, for several reasons. These include the following:

The DNA alphabet is smaller (4 letters), yielding less information for each
position (there are 20 possible amino acids at each position in a protein). The genetic code is redundant — there are several variations of most DNA codon triplets that code for an amino acid. Although a the protein product may be identical to your query sequence, you may not get an identical match with the DNA. Also, protein sequence similarity is more conserved through time than is DNA sequence similarity.

The search for protein orthologues is becoming increasingly important in molecular biology. Now that the complete *Saccharomyces cerevisiae* (yeast; a unicellular eukaryote) and *Caenorhabditis elegans* (nematode; a multicellular eukaryote) genomes have been sequenced, work is well underway to identify orthologous groups. If a novel human protein can be matched with orthologous proteins from yeast or The Worm, the investigator is likely to save a lot of time (and money!), having identified a likely function for the protein. [A good example is: Chervitz SA, et al (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*. 282:2022-2028.]

The Main Protein Sequence Databases

There are two major, non-specialised protein databases that you will frequently encounter: PIR and SWISS-PROT. Unlike the three major nucleotide databases, the entries in PIR and SWISS-PROT are not mirrored (copied). Each one has it's advantages and disadvantages, which you should consider before deciding which database to search.

Both databases are cross referenced with the nucleotide databases by the nucleotide database unique identifier (accession number; NID) or by PID, the Protein Identification Number which serves the same function.

PIR-International Protein Sequence Database

Previously called just PIR, this is the oldest molecular sequence database available (established 1984). The entries arise from international collaborative efforts and are organised biologically e.g. by structural, functional or evolutionary relationships. The entries include amino acid sequences, and in many cases further annotation including: citations (linked to Medline for abstracts); nucleotide database references; current genetic information (including map position and the start codon if not AUG).

*PIR is, in part, a redundant database.* Sequences are made public as soon as the database curators receive them, even before annotation or classification is verified. Redundancy has it's disadvantages, most notably the repetition of sequences in different entries may include discrepancies. The redundancy at PIR can be advantages, as sequences are made public very quickly. The database is updated weekly.

The PIR-International protein sequence database is partitioned into four sections: PIR1–PIR4. There is no clear cut difference between the entries in PIR1 and PIR2.

**PIR1**

Classified, annotated, verified and non-redundant with respect to other PIR1 entries.

**PIR2**

Essentially indistinguishable from PIR1. Classification may not be quite so extensive as in PIR1.

**PIR3**

Not classified, annotated or verified. No attempts have been made to reduce redundancy.

**PIR4**

Unencoded or untranslated.
SWISS-PROT Protein Sequence Database

SWISS-PROT (established 1986) is a protein sequence database, accessible from the Swiss EMBL Outstation, EXPASY. SWISS-PROT excels in annotation, exhibits very little redundancy and is thoroughly integrated with other databases. The extensive annotation and exhaustive to reduce redundancy mean that entries can take time before they are made available, but when they are, they are a complete and thorough resource.

Annotation is updated with information from published review articles, and by external expert referees. The entries are similar in layout to EMBL entries, with similar two letter codes defining the contents of each line. These include CC (comment), FT (feature table) and KW (keywords). Annotation includes information about the protein's function, post-translational modifications, disease associated deficiency, domains, structure and more.

Where applicable, SWISS-PROT entries are cross referenced with PDB, a database of experimentally determined protein structure. Three dimensional (3D) models can be viewed with most web browsers, or files can be downloaded for local viewing.

TrEMBL

TrEMBL is a supplement to SWISS-PROT that contains computer annotated translations of EMBL. When entry annotation and verification is complete, it is moved from TrEMBL to SWISS-PROT (assuming the entry does not already exist, in which case they will be merged). Since preparing entries for SWISS-PROT is so time consuming, TrEMBL basically attempts to bridge the gap, and provide a redundant database of (less extensively) annotated translations of coding sequences (CDS) that are not listed in SWISS-PROT.

TrEMBL has two main sections. SW-TrEMBL (SWISS-PROT TrEMBL), which contains sequences that are en route to SWISS-PROT. In contrast, REM-TrEMBL stores the remaining entries. This includes entries specifically excluded from SWISS-PROT, such as the many variations of immunoglobulins and T-cell receptors, synthetics sequences, fragments of less than eight amino acids, CDS from patent applications and EMBL CDS translations where the curators have strong evidence that the nucleotide does not code for real proteins.

Well done! You have completed the most important chapters in this guide!

SWISS-PROT: Exercise

You will recall BLAST and FASTA from the previous chapter. You can use these same programs to search protein sequence databases, as well as nucleotide sequence databases. Make sure you specify which database you want to search, and choose the right flavour of program to use (Hint: Blast programs and FASTA programs were described earlier).

In the exercise given below, you will integrate the knowledge you have gained from this guide. You should also realise how easy it is to use other databases and related sources of information, particularly now that you have an understanding of the molecular databases.

You will need to make your own decisions from the first instruction, and at times, things may not be clear until you do them. Do persist, and try to resolve any complications that arise by yourself. You will become more proficient if you do so. You might want to print this page before you start. You might also want to write down answers to questions, as you go along. The correct answers will be given on the next page. And finally, good luck!
1. Go to a sequence alignment program of your choice. You might choose to use:
   - The EXPASY **BLAST** server.
   - The GeneStream **FASTA** server.

2. Copy the following **human** amino acid sequence (given in the **one letter code**).

```
MSTAVLENPLGRLSDFGQETSYIEDCNQNGAISLIFSLKEE 
VGALAKLRLFEENDVNLTHIESRPSLRKDEYEFFFTHLKRSLPALTNIKILRHDI 
GATVHELSDRRKKDDTVFPPRTIQELDRFANQ discarding the same letters as those in the human sequence.
```

3. Paste the sequence into the query sequence window and adjust the options as necessary. You won't need to specify advanced options, but you should choose a program and database. For simplicity, please use the main SWISS-PROT database. You may wish to try other databases, but you should return to SWISS-PROT when continuing with this exercise.

4. Run the search and identify the protein. Use the link provided to see the SWISS-PROT report. If the link fails for any reason, you can do a text search of SWISS-PROT. Go to **SWISS-PROT** and search by the identifier you identified after the BLAST or FASTA search.

   When you have retrieved the SWISS-PROT report, view it with NICE-PROT. This presents the SWISS-PROT database entry in a more user-friendly manner, replacing two letter line codes with proper English descriptions, a colourful smart presentation and active links.

Well done! That was the most difficult part. Now, try to answer all of the questions below. You may need to look at pages that are linked from the SWISS-PROT report, but you not need to search further than the first page of any site. Answering all of the questions may take some time, but you will get a feel for what is available, and how to get it. You may even find yourself becoming fascinated by the report, and exploring on your own! Write down the answers, and see if you got them right by comparing your answers to the correct answers on the next page.

1. What is the SWISS-PROT name of the entry?
2. What is the SWISS-PROT primary accession number?
3. What is the most common name of the protein?
4. What is the gene called?
5. Which year was the crystal structure of the catalytic domain determined? Name the first name author.
6. Does the enzyme require a co-factor to function? If so, what?
7. Name the most common disease that arises as a result of deficiency of this enzyme.
8. Which cytogenetic locus does the gene reside at? (e.g. 13p10.1)
9. What is the PAHdb?
10. How many amino acid residues are there in the protein?
11. What is the molecular weight of the protein?
12. Tasks: Look briefly at entries in GeneCards, MIM (Mendelian Inheritance in Man), obtain the nucleic acid sequence and locate a FASTA report for the protein sequence. View a three-dimensional (3D) image of the protein that the gene codes for (Hint: PDB stores such files!). Finally, click on "HOMO SAPIENS (HUMAN)" and have a look at the taxonomy page.

The answers can be found on the next page, along with a brief discussion of...
how you should have found them, and how useful the information is!

**SWISS-PROT: Exercise Answers**

You could have chosen to use either BLAST or FASTA, but you would have been wise to use BLAST for the speed.

**If you used BLAST**, you should have chosen to use *blastp* and search the SWISS-PROT database (you did not need to add TrEMBL to the search). You should have identified human PAH1 as the correct sequence because it was the only *human* sequence with 100% identity. When you clicked on the identifier links, you should have been taken to the SWISS-PROT entry, where you would have clicked on *Nice Prot* to see the database entry in a nice layout.

**If you used FASTA**, you should have used the *fasta* program and searched the *swissprot S* database. You would have identified human PAH1 as the correct sequence because it was the only *human* sequence with 100% identity. You may not have been able to link directly to SWISS-PROT, in which case you would have searched SWISS-PROT directly using the keyword PH4H_HUMAN—the description from the FASTA report.

All of the following answers were found in the SWISS-PROT report, or on the first page of a link following it. For example, the gene locus is the first thing reported in the OMIM (Online Mendelian Inheritance in Man) report (link under heading MIM), and is also reported in the GeneCard for the enzyme. It may also have been listed elsewhere. Here are the results:

- **What is the SWISS-PROT name of the entry?**
  PH4H_Human

- **What is the SWISS-PROT primary accession number?**
  P00439

- **What is the name of the protein?**
  Phenylalanine-4-Hydroxylase

- **What is the gene name?**
  PAH

- **Which year was the crystal structure of the catalytic domain determined?**
  1997, Erlandsen

- **Does the enzyme require a co-factor to function? If so, what?**
  Yes. A ferrous ion.

- **Name the most common disease that arises as a result of deficiency of this enzyme.**
  Phenylketonuria (PKU).

- **Which cytogenetic locus does the gene reside at? (e.g. 13p10.1)**
  12q24.1

- **What is the PAHdb?**
  It's the Phenylalanine Hydroxylase Locus Database of mutations. A specialised database that concentrates purely on PAH. It includes comprehensive entries about PAH mutations.

- **How many amino acid residues are there in the protein?**
  452

- **What is the molecular weight of the protein?**
  51.862 kDa
Viewing the 3D molecule
You should have followed the link to the PDB database, under the heading of 'Cross References'. You should have been able to work out that you could either download and install software, or use the provided web-based viewer. All you had to do was click on the correct link in the PDB entry to see the 3D image!

I hope you did well. Since it's such a useful (and fun!) resource, the last page in this chapter will introduce you to PDB from where you can obtain and view 3D molecular structures.

The Brookhaven Protein Data Bank (PDB)

This database contains entries for molecular sequences, whose structure has been experimentally determined by X-ray crystallography or nucleic magnetic resonance imaging (NMR, MRI). I stress, the images presented have been experimentally acquired, and are not theoretical. Once you have located your desired entry, you can view the molecule in a number of ways. There are three common routes:

1. Download the "PDB file" for that entry. When you open the file using a three dimensional (3D) molecule viewing program (e.g. RasMol) on your computer, you will be able to interactively rotate, zoom, colour the molecule, and more.
2. Click on a link to see the two dimensional (2D) image of the molecule. This is the faster, more convenient method. You will see a 3D image in your web browser, but you will not be able to manipulate it in any way.
3. View the PDB file using a web based viewing program. This is more convenient than installing software on your computer, but may be slower to install, and may have less functions.

Clear, concise instructions are provided along with each PDB entry.

To try it, go to the PDB mirror at EBI, UK. From there, select PDB-Lite. You may like to try the non-lite version when you are familiar with the database. Enter the search term: Cystic Fibrosis and hit Return/Enter. The database should locate the CFTR molecule. Now decide whether you want to view the 3D image locally, in 2D or in a java applet in your web page. Follow the instructions on screen --they are very clearly presented.

You may like to visit the POV-RAY web site to download a free ray tracing program (available for most platforms), which you can use to render publication quality images of your PDB files. You may need a file format translator.

In case you cannot access the server, here are two examples of 3D images, displayed in two different ways:

1. CFTR Protein (95 K). Presented in space filling orbitals. Ray traced.
2. DNA molecule (62.5 K). Backbone of the molecule. Ray traced.
You have now completed the most important chapters in this guide! I congratulate you. If you have any comments or suggestions about this guide, please do not hesitate to send them to the webmaster. Other people like yourself will benefit from feedback, both positive and constructively critical.

The next chapter is about Molecular Sequence Alignment.

---

**Molecular Sequence Alignment**

**Introduction**

Welcome to the chapter on molecular sequence alignment. In this chapter, you will learn how to align sequences using freely available software on the WWW and from private servers. I will refer to nucleotides throughout this chapter, but you could just as easily align two protein sequences.

Sequence alignment is a useful tool with many, diverse applications. You might want to compare a new sequence against an established sequence that you have fetched from a molecular database, for example. On the other hand, you might be sequencing a new gene, in which case it is imperative that you sequence it accurately. One way to do this is to sequence it in both directions, then align the results (reversing one of them, of course!).

Sequence alignment involves metaphorically placing one sequence above the other and comparing the aligned vertical pairs at each position.
Two nucleotide or protein sequences can be compared against each other, quickly and accurately, by using computer software. Nucleotide sequences could be aligned manually, but automation speeds up the laborious process and eliminates the risk of human error. Aligning similar, nonidentical amino acid residues would be even more painstaking if done by hand.

Alignment Output Explained

Different programs use different ways of reporting matches, but a common method is to join a pair with a bar (|) to indicate that the sequences are identical. In this case, a colon (:) can be used to identify similar but nonidentical pairs. Such alignments involve IUB ambiguity codes e.g. N pairs with G, C, T or A. Nonidentical amino acids with similar physical properties can also be reported as similar. An example output (from GCG bestfit) is given below:

```
410 AANCGTGATCGATGCTAGCTATATA 434
 |||| |||| | | | | | | | | | |
410 AATCGTTATCGATGCTAGCTATATA 434
```

Let's dissect this example. I asked the alignment software to align two query sequences. Twenty five bases from each query have been reported in the alignment. The numbers at each end of the alignment corresponds to the nucleotide number in the original sequence, starting at one. We can therefore deduce that there was a 409 nucleotide nonidentical prefix in the query sequences. There may have been a nonidentical suffix too, or the entered sequence may only have been 434 bases long.

Looking at the aligned nucleotides, there are two nonidentical pairs. The first nonidentical pair (nucleotide number 413) is paired by a colon (:). The nucleotide at this position on the upper strand is an N indicating that the sequencer was unable to determine the nucleotide identity. If you require one base resolution, you should look at the sequencing gel (autoradiogram or electropherogram) again, to see if you can establish, without doubt, the identity of the nucleotide. Repeating sequencing is another option.

Nucleotide pair number 16 comprises a G and a T. These do not match and no line has been drawn between them. This appears to be the result of a point mutation.

FASTA Output

Let's take a brief look at another type of alignment output: the alignment reported after running a FASTA database search. An example is given below:

```
10 20 30
- TGCTAGACATGGATAATGGAGAGNCAATGTAAGTCC
 :::: :::: ::::: :::::::::
Z93016 TCTCTGATACAGTGTTGGAACAGATGA-TTGAGAGAAGACAATGGCTTCTA
 71780 71790 71800 71820 71830
```

The upper strand is the query strand, in this case named "-" as it was queried anonymously. The lower strand is a reported database sequence, found to match the query sequence. The first six letters of the sequence description are given on the left —in this case, a GenBank accession number.

Flanking nonidentical regions are reported to put the alignment into context. For FASTA reports, a colon (:) indicates homology (e.g. A with A), a period (.) indicates similarity (e.g. N with A) and an absence of a joining mark indicates that the pair of nucleotides are dissimilar (e.g. A
and G). We can also see an en dash (–) in the lower sequence. This indicates that a second A has been inserted into the query sequence. This is different to a mismatch because if this position is skipped (a – is inserted), the following nucleotides align perfectly.

Notice also the position of the alignment. The query sequence is reported from base 1 and aligns with the database sequence from base 71799.

Simple, isn't it? You should give it a try...

Alignment Using GCG Software

Warning! This page is only of interest to readers who have access to the Wisconsin GCG Package (referred to as GCG) on a UNIX server. If you don't have access, feel free to read on, but you should realise that you will not be able to use the licenced software.

GCG is a software package that can be run on a UNIX computer. You will usually run GCG on your remote server through a telnet session on your computer. If you don't understand this, contact your local user support or system administrator for help.

To complete the following exercises, you will need to be running GCG. If you are a registered user at the HGMP-Resource Centre, you may wish to telnet to menu.hgmp.mrc.ac.uk now. Use the menus and load the latest version of GCG. If you have an account elsewhere, you should open a telnet session now.

You should begin by loading the GCG package. At the UNIX prompt (your_username$), type:

$> gcg

The GCG Sequence Editor

You will need to enter your sequence into the GCG program called seqed, a sequence editor. If you are sequencing manually, you must type each base into the editor, one at a time. If you are using an automated sequencer with automatic base calling, you can copy the sequence and paste it into seqed.

The sequence editor will convert the nucleotide sequence into GCG format, required by GCG software and accepted by some WWW based packages too. (More frequently, FASTA format is required for web based software).

To start, you need to make a new file where the sequence will be stored. At the UNIX prompt in your telnet session, type the following command (using a new filename of your choice):

$> seqed filename

The seqed editor will now load. If it doesn't, you may have neglected to type gcg as instructed earlier, or you may have spelt seqed wrongly. Try again. A screenshot of seqed (size: 22.5Kb) might help if you get stuck, or if you don't have a telnet session open. At the top left, you can see the filename (newfile.seq in the screenshot). Your input cursor is below the filename, next to some colons. You can type a description of the sequence here i.e. what it is, where it came from and so forth.
When you have finished entering your description, hit D (i.e. on your keyboard, hold down the Control key as you hit D). The input cursor will move to the middle of the screen where you can paste or enter the sequence. Try it! It is normal practice to enter the nucleotides in capital letters (A, C...), and to use small letters (a, c...) to mark areas of special interest. Use N to designate a base whose identity is unclear.

When you have finished, hit D again to move the cursor to the command line at the bottom of the screen. Type ex and hit Enter/Return to save and exit seqed. To exit without saving, type q. For more commands, type help.

The saved file is now ready to be fed into GCG programs. To edit the file later, you can load seqed with the name of the file you just made:

```
$> seqed filename
```

To save time, you may wish to download an example sequence in GCG format. You can do so by saving the contents of the following link to disk: M13mp18.seq. This file is a text file UNIX line breaks. Don't open and save it in a non-UNIX compliant editor, or it will become unreadable by GCG!

**The GCG Alignment Software**

The bestfit program will determine the optimal alignment of two GCG format sequence files. The best region of similarity is reported (saved into a new text file). The Smith-Waterman algorithm is used (this is a good thing). A screenshot of bestfit (size: 27.0Kb) might help if you get stuck, or if you don't have a telnet session open.
To perform an alignment of two sequences saved in files one.seq and two.seq, you should type the following at the prompt:

$> \text{bestfit one.seq two.seq}

After a short introduction to the program, you are asked a number of questions. You can hit Enter/Return to accept the default parameters to all of these questions. The default variable is shown in brackets after each question. The first questions refer to the first sequence. When the questions are repeated, they refer to the second sequence. The questions are briefly explained below:

**Begin**

Enter the nucleotide or residue number that you wish to begin the alignment at. The first base in the sequence is counted as number one. The default is to start at the first.

**End**

Specify the last base in the sequence to be aligned. The default is to end at the last.

**Reverse**

The program can align the reverse of an input sequence. This is useful if you have sequenced the same region in both directions, for example, and wish to align the results.

**Gap Creation Penalty**

This is the penalty score to be awarded for the first nucleotide/residue in a gap. Gaps were discussed in the previous chapter of this guide.

**Gap Extension Penalty**

This is the penalty score to be awarded for each additional residue in a gap.

*If there are too many gaps or the gap is too long in a region of the sequence (i.e. the awarded penalty scores are too great), then that region will not be reported by bestfit.*

You should accept the default values, unless you have a reason to change them.

Finally, bestfit will propose a filename for the output (filename.pair) which you should accept, or enter an alternative. When bestfit has finished processing your sequences, you can view the output in your favourite text editor, or print it. For example:

To view: $> \text{pico filename.pair}

To print: $> \text{lpr printer_name filename.pair}.

The output was discussed on the previous page. At the top of the report, percentage similarity, refers to the similar bases including N pairing with any other base. The exact matches are reported as percentage identity.

You now know how to align sequences using GCG!

**Alignment Exercise Using WWW Software**

Basic sequence alignment is quite simple to perform. You will need to copy
and paste the given sequences one at a time, into the alignment program window. The instructions given below are simply guidelines. You should feel free to explore, although there aren't many options to choose between! You might like to try aligning polypeptide sequences.

Imagine that you have made mutant clones of a DNA sequence. How do you know which clones have been successfully mutated, and what the mutations are? You could sequence the clones, and align (compare) them with the original sequence! Have a go! The instructions are given below:

1. Go to the ALIGN server at GeneStream, France. (This program uses the BLOSUM50 matrix)
2. Copy the sequences given below and paste them into the query windows of the ALIGN program. Only paste the upper sequence into the upper input window, and the lower sequence into the lower window.

   First Sequence:
   AACGCCAGGGTTTTCCAGTCACGACGTTGTAAGCGACGGCCAGTGCCA

   Second Sequence:
   AACGCCAGGGTTTTCCAGTCACGACGTTGTAAGCGACGGCCAGTGCCA

3. Hit the "Perform Align" button.
4. Examine the output!

You will notice that matching pairs are aligned with a colon, mismatches have no connecting character and similar pairs (e.g. N and C) have a period joining them. Insertions and deletions are indicated by an en dash (–) in place of the missing nucleotide. You might like to try again, specifying different options, such as the xX marks. In this result format, no mark shows homology, X shows mismatch and x shows similar but not identical pairs.

Advanced users might like to browse the Pairwise Sequence Alignment program, which has several advanced features including the ability to align a DNA sequence with a protein or cDNA sequence. If aligning a protein sequence, take care when interpreting the output –START codons are not necessarily the first!

That concludes this section. Next, we will look at a variety of Further Sequence Analysis utilities.

---

Further Sequence Analysis

Introduction

STOP! An important note from the author. This chapter assumes previous knowledge. If you have not read: Molecular Databases, you should do so before reading any further.

In this section, you will be given an example to apply to several different programs. Your knowledge acquired so far should put you in good stead to understand what is going on, and enable you to use the programs for your own sequences.
The genetic code is known, and can be used to translate coding nucleotide sequences. For example, CCC codes for a proline amino acid, and GTA codes for valine. The translation software is non-selective, and will translate any nucleotide sequence that you provide. You should take care not to assume that the first protein sequence it returns, is the product of in vivo translation. The best way to check whether the translation is potentially a true one, is to align the returned protein sequence against a protein database using BLASTP for example.

**Exercise 1**

Some software packages will translate the nucleotide sequence in all three reading frames for each strand. To understand the difference, you should try The Protein Machine at EBI, UK. This server will translate your sequence in only one frame at a time. You must specify the reading frame, and whether you wish to align the complementary strand to the one you enter.

- Copy the nucleotide sequence below. It is the complete sequence for human exon 1 of the Lamin B₁ nuclear matrix attachment protein.

  GGGTGGGCACTCCAGCCGATGCTTCTCTCTTACAGCCCTGAGCTGGTCCGGGAACTTCCAGCCGGGAGGGCCGAGCTGACGGTTCGCAAGGGCCAGATTTTAAATTTACAGGCCCCGCCGGGGGAAGCCGCGCTGCTGCCGACAGCCTGCAAGGGCGCCGGCCAGATTTTAAATTTACAGGCCCGGCCCCCGAACGCCGAAGCGCGCTGCCGACCCCTACGCTGAGAGGGAAACAAAGTGCTGCGACGAGAGACGGCGGCGGCGCGAACCCTGCTGGGCCTCCAGTCACCCTCCTCTTGCATTTTCCCGCGTGCGTGTGTGAGTGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT
Finding Restriction Enzyme Sites

The **WebCutter** free, online program identifies all restriction enzyme cutting sites in your query sequence. The procedure is very easy: try and use the software yourself, using the sequence given above! You can specify the number of times your desired restriction enzyme should cut (an exact number or a range) and choose to only use restriction enzymes with a recognition site that is a certain number of bases long. You can choose to identify all of the restriction sites in the sequence, or look for a specific one. A screenshot of the sequence output (16 K) and the enzyme list (19 K) that appears at the bottom of the report, is available in case you cannot access WebCutter.

---

### Enzyme Sites Analysis

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Positions of Recognition Sites</th>
<th>Recognition Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc16I</td>
<td>803</td>
<td>tgc/gca</td>
</tr>
<tr>
<td>AccE7I</td>
<td>535</td>
<td>ccaannn/ntgg</td>
</tr>
<tr>
<td>AcsI</td>
<td>100</td>
<td>r/aatty</td>
</tr>
<tr>
<td>AcyI</td>
<td>17</td>
<td>gr/cgyo</td>
</tr>
<tr>
<td>AfaI</td>
<td>788</td>
<td>gt/ac</td>
</tr>
<tr>
<td>AlewNT</td>
<td>535</td>
<td>ccaannn/ntgg</td>
</tr>
</tbody>
</table>
Open Reading Frame Search

It is possible to feed a computer program a protein sequence, and ask it to identify all of the open reading frames in that sequence. **ORF Finder** is an example of such a program. You might like to use the sequence below to try it. The procedure is quite intuitive, and a screenshot (26.5 K) of the output is available; in the screenshot, the last reading frame appears to be the correct one, being the longest continuous one).

>gi|450395|gb|L27350|HUMHD01 Homo sapiens huntingtin (HD

Primer Selection

Other useful tools are the programs that find primers in your sequence, based on your criteria. If you are experienced with PCR, you will understand the options. If not, there is a description of each option at the bottom of the **Primer3** web page, served by the Norwegian EMBnet node. You might like to enter the sequence given below (a screenshot (33.5 K) of the results is available; notice the useful information provided including G:C content and recommended annealing temperature (tm)):  

>gi|450395|gb|L27350|HUMHD01 Homo sapiens huntingtin (HD

GGCTGAGCGACGACGCAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA
GACTGTGGCTGCGCTGGACCTGAGGGAAACCTGCCTGTACGTGAGGCCCTAAAAAGCCAGAGACCTCACTCC
CGGGGAGCCAGCATGTCCACTGCGGTCCTGGAAAACCCAGGCTTGGGCAGGAAACTCTCTGACTTTGGAC
AGGAAACAAGCTATATTGAAGACAACTGCAATCAAAATGGTGCCATATCACTGATCTTCTCACTCAAAGA
AGAAGTTGGTGCATTGGCCAAAGTATTGCGCTTATTTGAGGAGAATGATGTAAACCTGACCCACATTGAA
TCTAGACCTTCTCGTTTAAAGAAAGATGAGTATGAATTTTTCACCCATTTGGATAAACGTAGCCTGCCTG
CTCTGACAAACATCATCAAGATCTTGAGGCATGACATTGGTGCCACTGTCCATGAGCTTTCACGAGATAA

Protein Motif Searching

Not only can you search databases for homologous sequences, but you can also search for protein motifs that have been conserved through evolution. If you have identified a new protein and you don't know anything about it, you would find it useful to perform a motif search to see if any other sequences have a homologous motif. The results might give you a clue as to its function. Alternatively, you might wish to perform a motif search of your mutated sequence to see if an important motif such as a transporter signal sequence has been corrupted.

Try running a search using the PRINTS server at EBI, UK. Paste in the complete sequence and hit the button "Run PPSRCH". A screenshot (20 K) of the output is available. The ABC transporter family of carrier proteins (recognised in this sequence) includes the multidrug resistance ATPase in mammalian cells, chloroquine-resistance ATPase in Plasmodium falciparum and many others. In the CFTR protein, the ABC transporter is responsible for carrying the chlorine channel out of the endoplasmic reticulum, and into the plasma membrane. A single amino acid mutation (ΔF508) is responsible for the protein accumulating in the ER, and consequently the onset of disease.
Conclusion

You have now completed this guide. If you have any comments, suggestions or even corrections (!), please do not hesitate to e-mail them to the webmaster.

End of Guide.

Appendix

Glossary

This page provides a brief glossary of terms and abbreviations commonly
encountered in bioinformatics. This glossary forms part of an online Guide to Molecular Sequence Analysis.

Some of these explanations are rather simplistic, in favour of brevity. Please refer to molecular biology text books for more comprehensive details.

Alu
A family of approx. 300 bp repetitive sequences, found dispersed throughout the human genome. Almost any 100 kb human nucleotide sequence will have Alu sequences within it.

Base Analogue
A chemical compound which is sufficiently similar to one of the nitrogenous bases normally found in DNA, that it can replace it. Base analogues may cause mutations, or be used in a modified PCR reaction (e.g. when sequencing).

Bioinformatics
The discipline of obtaining information about genomic or protein sequence data. This may involve similarity searches of databases, comparing your unidentified sequence to the sequences in a database, or making predictions about the sequence based on current knowledge of similar sequences. Databases are frequently made publically available through the Internet, or locally at your institution.

BLAST
A set of programs, used to perform fast similarity searches. Nucleotide sequences can be compared with nucleotide sequences in a database using BLASTN, for example. Complex statistics are applied to judge the significance of each match. Reported sequences may be homologous to, or related to the query sequence. The BLASTP program is used to search a protein database for a match against a query protein sequence. There are several other flavours of BLAST.

BLAST2
A newer release of BLAST. Allows for insertions or deletions in the sequences being aligned. Gapped alignments may be more biologically significant.

cDNA
Complementary DNA. DNA copies of the mRNA expressed in a specified tissue. cDNA sequencing has the advantage of only representing expressed genes. Since only ~3% of the vast quantity of DNA in the human genome are coding sequences, cDNA sequencing is particularly useful in certain situations. See EST.

CDS or cds
Coding sequence.

Clone
Population of identical cells or molecules (e.g. DNA), derived from a single ancestor.

Cloning Vector
A molecule that carries a foreign gene into a host, and allows/facilitates the multiplication of that gene in a host. When sequencing a gene that has been cloned using a cloning vector (rather than by PCR), care should be taken not to include the cloning vector sequence when performing similarity searches. Plasmids, cosmids, phagemids, YACs and PACs are example types of cloning vectors.

Consensus Sequence
A derived nucleotide sequence that represents a family of similar sequences. Each base in the consensus sequence corresponds to the base
most frequently occurring at that position, in the real sequences.

Contig
A DNA sequence that overlaps with another contig. The full set of overlapping sequences (contigs) can be put together to obtain the sequence for a long region of DNA that cannot be sequenced in one run in a sequencing assay. Important in genetic mapping at the molecular level.

DNA Sequencing
The experimental process of determining the nucleotide sequence of a region of DNA. This is done by labelling each nucleotide (A, C, G or T) with either a radioactive or fluorescent marker which identifies it. There are several methods of applying this technology, each with their advantages and disadvantages. For more information, refer to a current text book. High throughput laboratories frequently use automated sequencers, which are capable of rapidly reading large numbers of templates. Sometimes, the sequences may be generated more quickly than they can be characterised.

Downstream
Toward the 3' end of a nucleotide sequence.

EMBL
European Molecular Biology Laboratories. Maintain the EMBL database, one of the major public sequence databases.

EMBnet
European Molecular Biology Network: http://www.embnet.org was established in 1988, and provides services including local molecular databases and software for molecular biologists in Europe. There are several large outposts of EMBnet, including EXPASY.

EST
See Expressed Sequence Tag

Exon
Coding region of DNA. See CDS.

Expressed Sequence Tag (EST)
Randomly selected, partial cDNA sequence; represents its corresponding mRNA. dbEST is a large database of ESTs at GenBank, NCBI.

HGMP
Human Genome Mapping Project.
The UK HGMP Resource Centre is an academic institution in the UK which provides a number of services, including access to databases, mirrors of databases, and access to extensive services/software for registered academic users.

Intron
Non-coding region of DNA.

MMDB
Molecular Modelling Database. A taxonomy assigned database of PDB (see PDB) files, and related information.

NCBI
National Center for Biotechnology Information (USA). Created by the United States Congress in 1988, to develop information systems to support the biological research community.

NIH
National Institutes of Health (USA).
OMIM
Online Mendelian Inheritance in Man. Database of genetic diseases with references to molecular medicine, cell biology, biochemistry and clinical details of the diseases.

ORF
Open Reading Frame. A series of codons (base triplets) which can be translated into a protein. There are six potential reading frames of an unidentified sequence; TBLASTN (see BLAST) translates a nucleotide sequence in all six reading frames, into a protein, then attempts to align the results to sequences in a protein database, returning the results as a nucleotide sequence. The most likely reading frame can be identified using on-line software (e.g. ORF Finder).

Orthologue
Groups of genes or proteins from different organisms that have the same function, are said to be orthologous. There are numerous genes that have been conserved through evolutionary history. The protein products can be identified in yeast, a nematode worm and human cells, for example. It can be interesting to study the gene function in a worm, if you know that it has the same function in humans.

PDB
Brookhaven Protein Data Bank. A database and format of files which describe the 3D structure of a protein or nucleic acid, as determined by X-ray crystallography or nuclear magnetic resonance (NMR) imaging. The molecules described by the files are usually viewed locally by dedicated software, but can sometimes be visualised on the world wide web.

PIR
A database of translated GenBank nucleotide sequences. PIR is a redundant (see Redundancy) protein sequence database. The database is divided into four categories:

1. PIR1 - Classified and annotated.
2. PIR2 - Annotated.
3. PIR3 - Unverified.
4. PIR4 - Unencoded or untranslated.

Redundancy
The presence of more than one identical item represents redundancy. In bioinformatics, the term is used with reference to the sequences in a sequence database. If a database is described as being redundant, more than one identical (redundant) sequence may be found. If the database is said to be non-redundant (nr), the database managers have attempted to reduce the redundancy.

The term is ambiguous with reference to genetics, and as such, the degree of non-redundancy varies according to the database manager's interpretation of the term. One can argue whether or not two alleles of a locus defines the limit of redundancy, or whether the same locus in different, closely related organisms constitutes redundancy. Non-redundant databases are, in some ways, superior, but are less complete. These factors should be taken into consideration when selecting a database to search.

Sequence Tagged Site
Short cDNA sequences of regions that have been physically mapped. STSs provide unique landmarks, or identifiers, throughout the genome. Useful as a framework for further sequencing.

STS
See Sequence Tagged Site

SWISS-PROT
A non-redundant (See Redundancy) protein sequence database. Thoroughly
annotated and cross referenced. A subdivision is TrEMBL.

**TrEMBL**
A protein sequence database of Translated EMBL nucleotide sequences.

**UniGene**
Database of unique human genes, at NCBI. Entries are selected by near
identical presence in GenBank and dbEST databases. The clusters of
sequences produced are considered to represent a single gene.

**Upstream**
Toward the 5' end of a nucleotide sequence.

---

**FASTA Format Explained**

The genomic sequence below is in FASTA format, which is often required when
searching molecular databases. **Take care!** The first line **must** begin with
'>' and a short description! The description can be anything that you choose
to write. If you obtained the FASTA report from GenBank or EMBL, the
description can be read as follows:

**GenBank generated FASTA report:**
   >gb|accession|locus|description

**EMBL generated FASTA report:**
   >emb|accession|locus|description

**DDBJ generated FASTA report:**
   >dbj|accession|locus|description

**SWISS-PROT generated FASTA report:**
   >sp|accession|entry name

**nr (non-redundant) database generated FASTA report; sequences derived from
other databases:**
   >gi|gi_identifier|accession of nucleotide sequence from which it was
derived|description

*Accession* and *locus* refer to the ACCESSION and LOCUS numbers in the
database.

The DNA sequence of exon 11 coding for human lamin B₁ is given below. Lamin
B₁ is a protein found in the nucleus, that helps to organise the chromatin
during interphase. It does so by binding both to the nuclear matrix (a
fibrous scaffold just below the inner nuclear envelope) and special regions
of DNA. When you are confident, you might like to use this sequence to try a
specialised database search to identify matrix attachment regions (MARs) of
a DNA sequence, of which Lamin B is a classic example. The [TimeLogic Server](#) requires registration, but offers the MAR Finder service.

Sequences reported as being mRNA are in fact cDNA, obtained from mRNA. You
will notice the absence or uracil (only present in RNA) in such sequences.

*Hit the BACK button on your web browser, to return to the previous page.*
Reference List

Database Reviews

1. **GenBank:**

2. **EMBL:**

3. **DDBJ:**

4. **PIR:**


BLAST


**FASTA**


**Matrices**


**Match Scoring**


**Filtering Programs**


2. **dust**: Tatusov and Lipman.


**Other Interest: ESTs**


Other Interest: Books


Published April 1999.

**Author: Andrew Louka**
Department of Biology & Biochemistry
Brunel University
Uxbridge
Middlesex
United Kingdom.
E-mail: andrew.louka@embnet.uio.no or check [http://www.whowhere.com/](http://www.whowhere.com/)