3

Sequence Comparisons and Sequence-Based Database Searches

P.M. Selzer et al., *Applied Bioinformatics*, https://doi.org/10.1007/978-3-319-68301-0_3

3.1 Pairwise and Multiple Sequence Comparisons

The comparison of protein and DNA sequences is an important analytical method of applied bioinformatics. The annotations of new nucleotide and protein sequences, construction of model structures for proteins, design and analysis of expression studies, and a variety of other bioinformatic and biological experiments are all based on these analyses. Nature acts conservatively, i.e., it does not develop a new kind of biology for every life form but continuously changes and adapts a proven general concept. Novel functionalities do not appear because a new gene has suddenly arisen but are developed and modified during evolution. Given this situation, therefore, one may transfer functional information from one protein to another if both possess a certain degree of similarity. However, this process must be carried out critically because similar proteins may yet perform different functions. The similarity of two proteins can arise based on evolution from a common ancestor (convergent evolution) or independently of each other based on different ancestor proteins (divergent evolution).

Before analyzing whether sequences are possibly related, it is first necessary to define some terms. Related sequences are designated as being homologous; however, the term *homology* often leads to confusion. Homology is not a measure of similarity, but rather signifies that sequences have a shared evolutionary history and, therefore, possess a common ancestral sequence (Tatusov et al. 1997). The definitions of the terms *ortholog* and *paralog* in combination with the function of a protein are, however, the subject of controversy (Jensen 2001; Gerlt and Babbitt 2001). In general, genome biologists define these terms as follows. Homologous proteins from different species that possess the same function (e.g., corresponding kinases in a signal transduction pathway in humans and mice) are called orthologs. In contrast, homologous proteins that have different functions in the same species (e.g., two kinases in different signal transduction pathways of humans) are termed paralogs.

Homology is not quantifiable – either two sequences are homologous or they are not. The identity or similarity of two sequences is, however, quantifiable. Identity is the ratio of the number of identical amino acids or nucleotides in a sequence to the total number of amino acids or nucleotides. Unlike identity, similarity is not as simple to calculate. Before similarity can be determined, the degree of similarity of the building blocks of sequences to each other must first be determined. This is done using similarity matrices that are also known as substitution or scoring matrices. Similarity matrices specify the probability that a sequence will transform into another sequence over time. Of course, this depends on the time elapsed and the mutational rate of nucleotides. Identity is an absolute measure that is, in contrast to similarity, not based on a specific defined model or a similarity matrix. Sequences of two homologous proteins can have a similarity of 60% and an identity of 40%. They do not show a 60% or 40% homology, a statement that can sometimes even be found in standard literature. It should be noted that similarity can be used only for amino acid sequences and not nucleotide sequences.

Before deciding upon the identity or similarity of two nucleotide or amino acid sequences, an alignment must first be calculated. The underlying principle of such an alignment is relatively simple (Fig. 3.1). Two sequences are arbitrarily placed next to each other and the alignment judged according to the quality measure (e.g., a similarity matrix). The two sequences are then moved relative to one other, and for each position a score is calculated. This process is repeated until the best alignment is found.

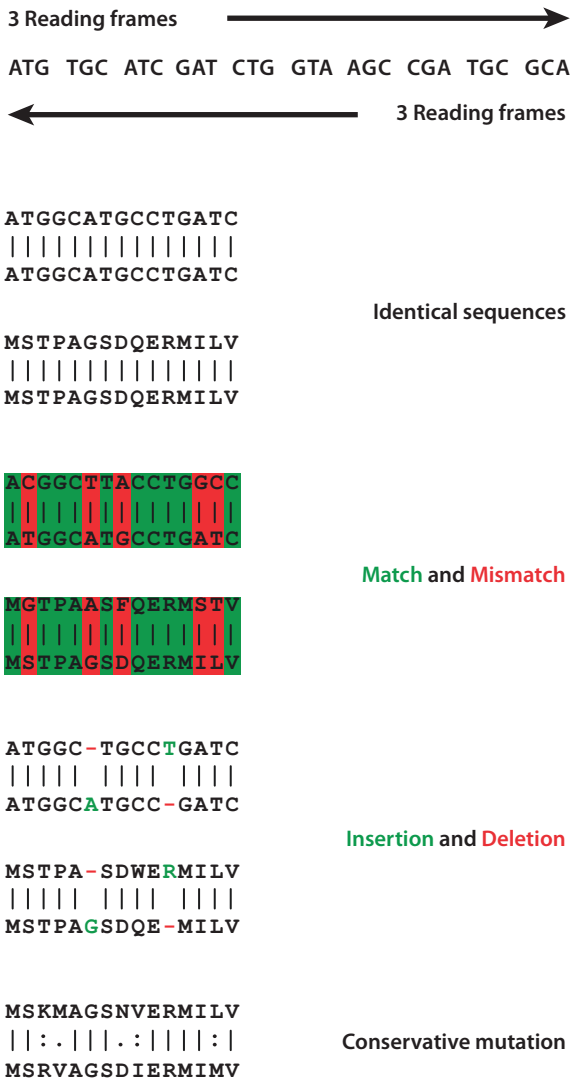
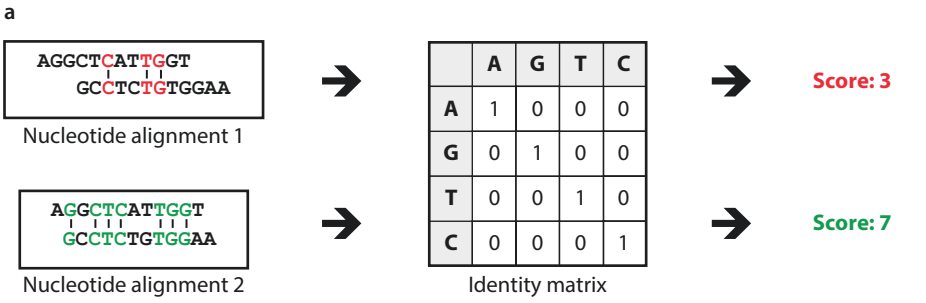


Fig. 3.1 Sequence alignments of nucleotide and amino acid sequences

The determination of the quality measure is the real challenge for this process. For nucleotide sequences the simplest solution is an identity matrix (Fig. 3.2a). Here, one assumes that the four nucleotides do not show any similarity to one other, and therefore, only identical nucleotides are factored into the similarity scoring. They are regarded as identical (match) or different (mismatch). For the final scoring only identical nucleotides are added up.

For protein sequences, an identity matrix is not sufficient to describe biological and evolutionary processes. Amino acids are not exchanged with the same probability as might be conceived theoretically. For example, an exchange of aspartic acid for glutamic acid is frequently observed; however, a change from aspartic acid to tryptophan is rarely seen. One reason for this is the triplet-based genetic code (▶ Chap. 1). For an exchange



b

GMIDLITARCAYPSWTGH
IEVRTAKCAYPGWSGHY
Amino acid alignment 1

GMIDLITARCAYPSWTGH
IEVRTAKCAYPGWSGHY
Amino acid alignment 1

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	-2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-3	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

BLOSUM62 matrix

Score: 65

Score: 19

Fig. 3.2 Scoring matrices allow the computation of optimal alignments. **a** Use of an identity matrix for the construction of an optimal nucleotide alignment. **b** Use of the BLOSUM62 matrix for the construction of an optimal amino acid alignment. Two potential alignments for each are represented; the optimal alignment is shown in green

of aspartic acid to glutamic acid to occur only a mutation of the last nucleotide in the triplet codon is required (GAT/GAC to GAA/GAG). In contrast, a complete mutation of the whole triplet must occur in order for aspartic acid to be exchanged for tryptophan (GAT/GAC to TGG). Of course, such a complete mutational substitution has a much

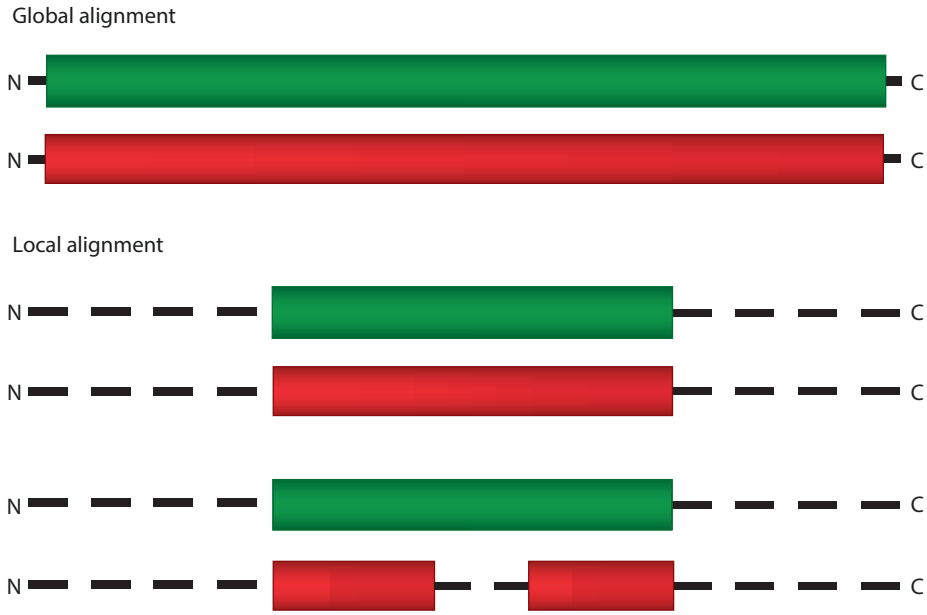


Fig. 3.3 Global and local sequence alignment. Gaps can also appear in a global alignment, as seen in the lower local alignment

lower probability of occurring and needs a longer timeframe. A second reason for the mutation of aspartic acid to glutamic acid to occur more often is that both have similar properties (▶ Chap. 1). In contrast, aspartic acid and tryptophan are chemically different – the hydrophobic tryptophan is frequently found in the center of proteins, whereas the hydrophilic aspartic acid occurs more often at the surface. An exchange of aspartic acid for tryptophan, therefore, could greatly alter the tertiary structure of a protein and, consequently, its function. Such striking amino acid exchanges accompanied by a loss of function rarely happen.

Therefore, most algorithms use substitution matrices to align protein sequences. These amino acid substitution matrices describe the probability that amino acids will be exchanged in the course of evolution. These matrices contain a logarithm for the relationship of two probabilities that a couple of amino acids or nucleotides will appear in an alignment, i.e., both the probability of a coincidental concurrence and the probability of an evolutionary event responsible for the occurrence are taken into account. Negative values in the matrix mean that the occurrence is rather coincidental, whereas positive values suggest an evolutionary event. Because the matrix values are logarithms of relationships, addition of the numbers leads to a conclusion for the complete alignment. The most commonly used amino acid scoring matrices are the position accepted mutation (PAM) (Dayhoff et al. 1978) and blocks substitution matrix (BLOSUM) groups (Henikoff and Henikoff 1992) (Fig. 3.2b).

Alignments can be carried out both globally and locally (Fig. 3.3). In global alignments, complete nucleotide or protein sequences are compared to one another over the

a

	M	T	P	A	R	G	S	A	L	S
M	5	-1	-2	-1	-1	-3	-1	-1	2	-1
T	-1	5	-1	0	-1	-2	-1	0	-1	-1
P	-2	-1	7	-1	-2	-2	-1	-1	-3	-1
V	1	0	-2	0	-3	-3	-2	0	-1	-2
R	-1	-1	-2	-1	5	-2	-1	-1	-2	-1
R	-1	-1	-2	-1	5	-2	-1	-1	-2	-1
S	-1	1	-1	1	-1	0	4	1	-2	4
L	2	-1	-3	-1	-2	-4	-2	-1	4	-2
S	-1	-1	-1	-1	-1	0	4	1	-2	4

b

	M	T	P	A	R	G	S	A	L	S
M	32									
T		27								
P			22							
V				15						
R					15					
R						10				
S							12			
L								8	8	
S										4

c

```
1 MTPARGSALS 10          Length: 10
  |||.|.| ||            BlOSUM62, Gap_penalty: 1.0
1 MTPVRRS-LS  9          Score: (32.0-1.0) = 31.0
```

Fig. 3.4 Calculation of a global alignment of two similar protein sequences. **a** Both sequences are compared in a two-dimensional matrix, and the similarity of the amino acids is determined using similarity matrices. Each alignment can be described as a path through the two-dimensional matrix, starting with the highest-scoring amino acid pair at the N-terminus. **b** Adding the values produces corresponding scores for the different paths. The alignment with the highest score is considered optimal (red). **c** The optimal alignment is obtained by the introduction of a gap and contains 10 amino acids, of which 7 are identical. Using the BLOSUM62 similarity matrix and a gap penalty of 1.0 a score of 31.0 is achieved

entire length of the sequence. Figure 3.4 shows the calculation of a global alignment. However, even very similar sequences can have single deletions or insertions and, consequently, a different number of amino acids or nucleotides. To represent these alignments appropriately, gaps must be inserted into the sequences. Theoretically all possible sequences can be aligned by the introduction of gaps. To prevent this, fixed values (the scoring penalties) are given for the introduction of gaps (gap opening) and their extension (gap extension). These penalties are then subtracted from the alignment score to yield the total score. The alignment with the highest total score is considered the optimal sequence comparison. This method is based on the algorithm of Needleman and Wunsch (1970).

Sometimes, interest may focus solely on aligning the most similar stretches within two sequences – a local alignment (Fig. 3.3). This approach makes it possible to identify protein domains and motifs (e.g., ATP binding sites, DNA binding domains, N-glycosylation sites). In principle, a local alignment is calculated in the same way as a global alignment using a substitution matrix and the introduction and extension of gaps. The difference lies in the score calculation and the path through the matrix. A negative score is replaced by a zero, and so the path through the matrix does not move from the lower right to the upper left but starts and ends at arbitrary places (Fig. 3.4). The local

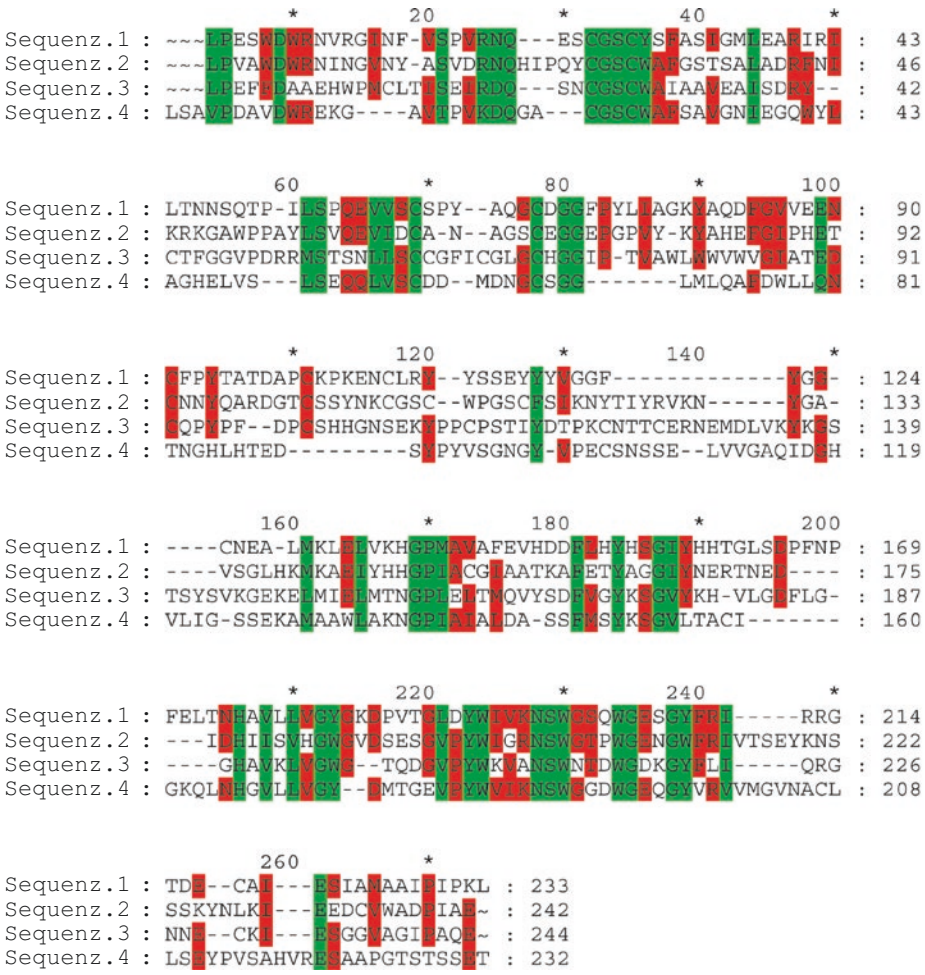


Fig. 3.5 Multiple-sequence alignment of four related proteins. Amino acids conserved in all four sequences (or with conservative changes) are highlighted in green; those conserved only in three of four sequences are shaded red

alignment identified in the matrix with the highest score is regarded as optimal and the starting point. The alignment ends when a zero entry is reached. This method is based on the algorithm of Smith and Waterman (1981).

To compare more than two nucleotide or protein sequences, one could compare all sequences pairwise and then further examine these alignments. However, it is quicker to perform a multiple alignment (Fig. 3.5) and analyze the overall alignment. One well-known program is ClustalW (Thompson et al. 1994), which was subsequently superseded by Clustal Omega [clustalomega] (Sievers et al. 2011). These programs utilize the fact that similar sequences are usually homologous. The basis for multiple sequence alignments are the pairwise alignments of all sequences. A phylogeny tree that represents the evolutionary relationship between the sequences in a tree structure

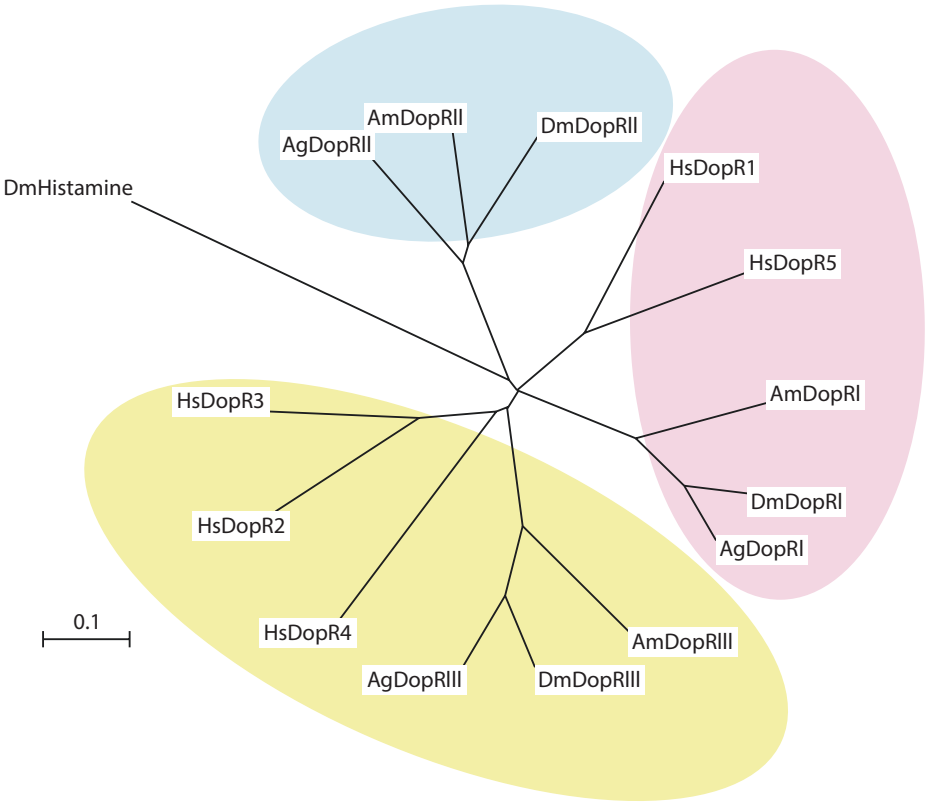


Fig. 3.6 Phylogenetic tree of dopamine receptor sequences. The evolutionary relationship between the sequences is reflected by the length of the branches. Dopamine receptor sequences of invertebrates (*Dm*, *Drosophila melanogaster*; *Ag*, *Anopheles gambiae*; *Am*, *Apis mellifera*) are compared with those of humans (*Hs*, *Homo sapiens*). Three clear clusters are formed. As a control, the phylogenetically distant sequence of the *Dm* histamine receptor was not found in any of the clusters

is then constructed. The evolutionary distances correspond to the length of the horizontal branches (Fig. 3.6). The final multiple-sequence alignment starts with the two most similar sequences. Step by step, the next most similar sequence is then added and aligned until the final multiple alignment is retrieved.

3.2 Database Searches with Nucleotide and Protein Sequences

A frequently used application of pairwise alignments is the search for similar protein or nucleotide sequences in sequence databases. With older dynamic alignment algorithms such as those designed by Smith and Watermann (1981) or Needleman and Wunsch (1970), this is too slow to perform even on current computers. Instead, heuristic algorithms like Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990; Boratyn et al. 2013) are employed (Figs. 3.7 and 3.8). Heuristic methods make assessments

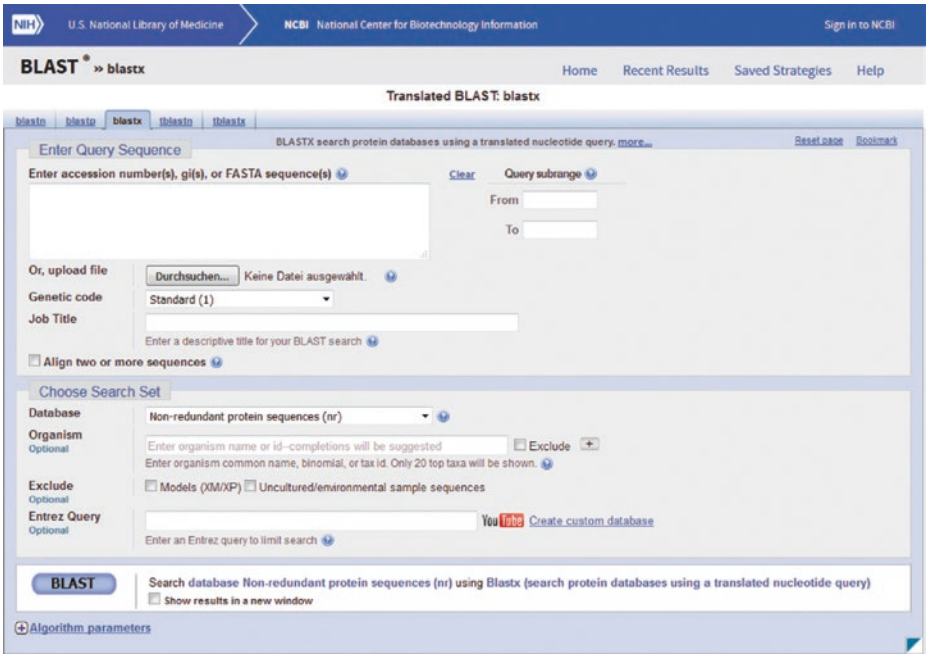


Fig. 3.7 Translated BLAST start page at NCBI. The blastx algorithm was used to compare a nucleotide sequence with a protein sequence database (Printed with permission of NCBI)

to obtain almost exact results and utilize sequence and alignment statistics to make searches in large databases feasible. They do not guarantee an optimal alignment, however, but allow for sensitive and fast database searches. BLAST is provided as a Web service at the NCBI [ncbi-blast], the EMBL-EBI [embl-blast], and the DDBJ [ddbj-blast]. BLAST is usually performed first against a nonredundant database, which is a compilation of entries from different databases. In a nonredundant database, multiple entries are removed so that every record is available only once. These databases exist for both nucleotide and protein sequences.

To execute a meaningful search in a nucleotide or protein database, the corresponding algorithm must be chosen from the BLAST group, and this depends on the aim of the search as well as the nature of the query sequence (nucleotide or protein) (Table 3.1). For example, to query a nucleotide database with a protein sequence, every nucleotide sequence of the database must be translated into all six theoretically possible protein sequences since the reading direction and the triplet starting point is unknown (Fig. 3.1). Only then can the query sequence be compared with the database. This complex process is performed automatically by the algorithm tblastn. Depending on the nature of the query and the databases used, a total of five algorithms are possible (Table 3.1).

In cases of a very low sequence identity, apparently meaningful but random alignments could be found that are not based on a common ancestor sequence. Therefore, the E-value is part of every BLAST result and can be used to assess the significance of an alignment. Pairwise alignments with an E-value < 0.02 are regarded as meaningful

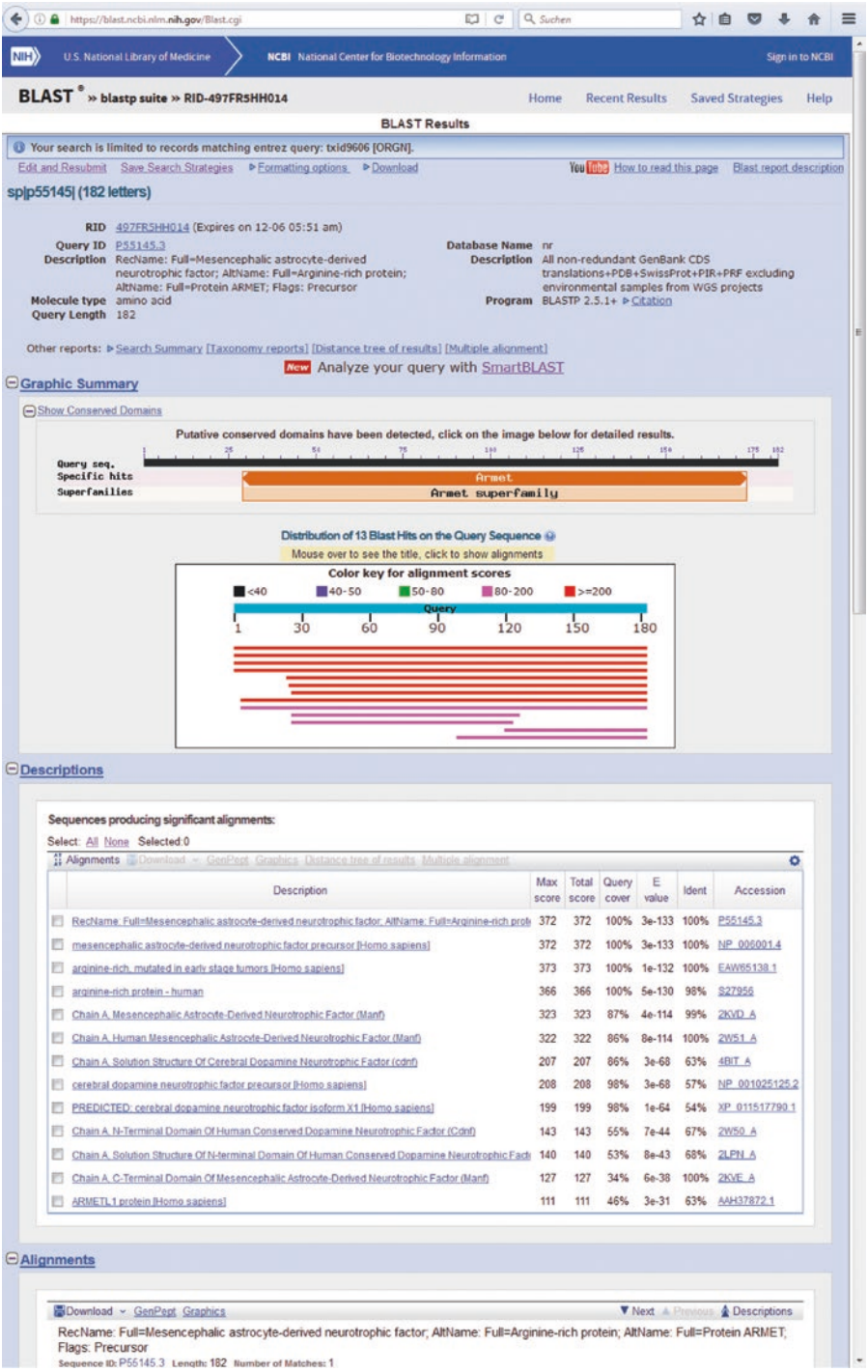


Fig. 3.8 Graphic representation of a BLAST result. The graph summarizes the number and length of hits with respect to the query sequence. The quality (alignment score) of the hits is represented by color coding (Printed with permission of NCBI)

Table 3.1 The most important algorithms of the BLAST group and their applications

Algorithm	Query sequence	Database	Remarks
blastp	Protein	Protein	–
blastn	Nucleotides	Nucleotides	–
blastx	Nucleotides	Protein	Query sequence is translated into all six reading frames
tblastn	Protein	Nucleotides	Database is translated into all six reading frames
tblastx	Nucleotides	Nucleotides	Query sequence and database are translated into all six reading frames

based on homologous sequences. A random alignment shows E-values > 1. The range in between needs further information (e.g., similar function) for a statement about homology.

Within the BLAST family of algorithms, Position-Specific Iterated BLAST (PSI-BLAST) (Altschul et al. 1997), Pattern-Hit Initiated BLAST (PHI-BLAST) (Zhang et al. 1998), and bl2seq (blast two sequences) (Tatusova and Madden 1999) are particularly interesting. The bl2seq algorithm carries out a local alignment of two sequences. PHI-BLAST allows searching for proteins in a protein database with sequence motifs similar to those of the query. PSI-BLAST is a mixture of a pairwise and a multiple alignment. First, a normal BLAST search is executed. With the resulting multiple alignment of hits, a sequence profile is constructed, which is then used to continue the search for new sequences until no more are found. The interpretation of the results is frequently very difficult and occasionally misleading because sequences not directly related can also be taken into account. Therefore, PSI-BLAST results require careful examination. Hidden Markov models (HMMs) (Eddy 2004) operate in a similar fashion, but more slowly and with greater sensitivity. Again, results from HMMs must be checked critically. The Conserved Domains Search by NCBI recognizes conserved domains within the analyzed sequences (Marchler-Bauer et al. 2015). There are also a number of species-specific BLAST applications for human, microbial, and other genomes as well as for the analysis of expression or immunological data and other special cases. These are available at the NCBI-BLAST Web page [ncbi-blast].

3.2.1 Important Algorithms for Database Searching

Needleman and Wunsch (1970) A global alignment that was first developed without gap functionality: The method uses a dynamic procedure, which is more efficient and faster in comparison with the calculation of all possible alignments. This calculation is still too time-consuming for the analysis of huge databases. It is very time intensive due

to its dynamic procedure. A dynamic procedure is a solution to a problem that is broken down into subproblems, and the best results are then compared.

Smith and Waterman (1981) A local alignment that was originally developed without gap functionality: The method is very similar to that of Needleman and Wunsch and also quite time-consuming.

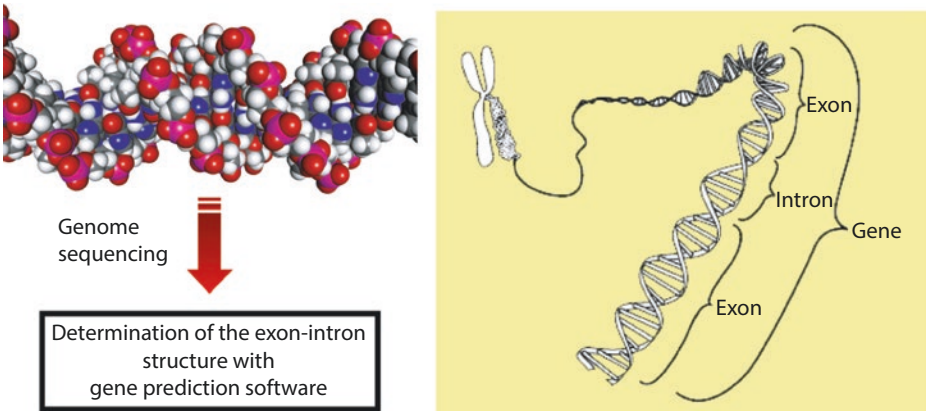
FastA (Pearson and Lipman 1988) A local alignment that is very fast due to the use of a heuristic method (making assessments to get almost exact results): The method identifies short word regions and then uses a dynamic procedure to obtain a gapped alignment.

BLAST (Altschul et al. 1990) A local alignment that can identify segment pairs of fixed length quickly due to the use of a heuristic method. Segments are then prolonged until preset threshold parameters are reached. BLAST is up to 100-fold faster than the Smith and Waterman algorithm.

Gapped BLAST (Altschul et al. 1997) A local alignment that looks only for a single segment pair: This segment pair is then prolonged by gaps in both directions. The gapped BLAST algorithm is three times faster than the ungapped BLAST algorithm.

3.3 Software for Sequence Analysis

Besides gene and protein sequences, NCBI, EBI, and other publicly accessible servers also provide genomic sequences. Such sequences are usually raw since they are published directly by sequencing units such as the Sanger Institute [sanger]. The advantage of raw sequence data is that predicted genes can be directly identified (◘ Fig. 3.9). A number of software solutions for gene predictions are offered on the Web. The Genscan server



◘ **Fig. 3.9** Identification of new genes and proteins by genome sequencing

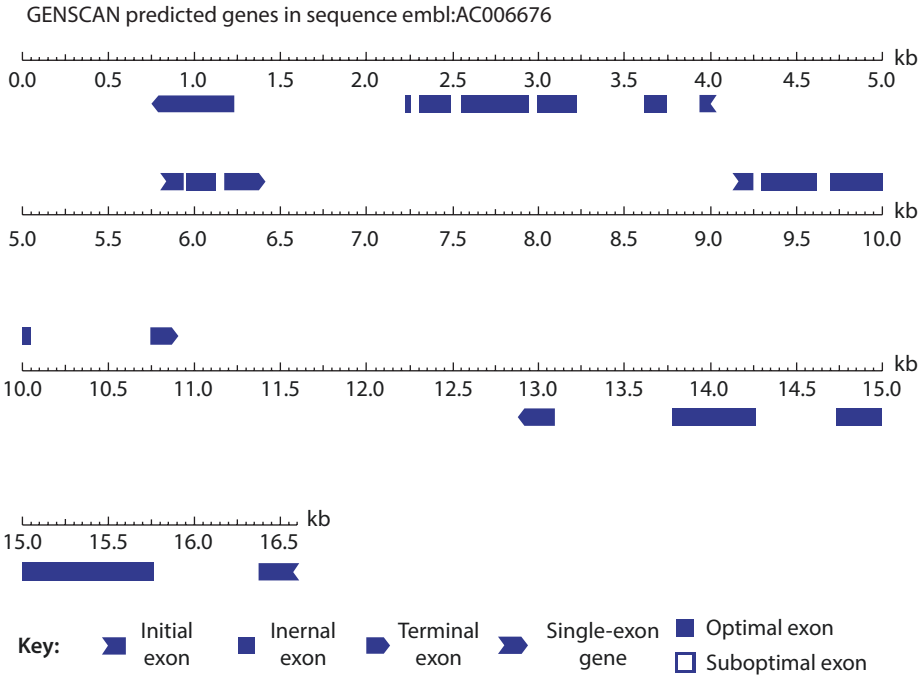


Fig. 3.10 Graphic version of result of a Genscan analysis [genscan]

of the Massachusetts Institute of Technology [genscan] is particularly important and is based on neural networks that are trained to extract the exon–intron structure of eukaryotic genes from genomic sequences. A typical result of a Genscan analysis is shown in Fig. 3.10. Another software available for gene prediction in prokaryotic sequences is Glimmer from the TIGR institutes (now part of the J. Craig Venter Institute), which is available in its third version at the Center for Computational Biology of John Hopkins University [glimmer].

An interesting development in the area of sequence analysis is the European Molecular Biology Open Software Suite (EMBOSS) (Rice et al. 2000) [emboss]. EMBOSS is an open-source project for different UNIX and Linux operating systems. The functional range of the software package grows steadily and is comparable with commercial packages such as that from GCG Wisconsin (Biova), the DNA-Star (DNASTAR Inc.), or Vector NTI software (Thermo Fischer Scientific Inc.). Expasy [expasy] and EMBnet [embnet] should also be mentioned. Besides databases, Expasy offers a number of hyperlinks to bioinformatic software. EMBnet is a worldwide association of different research groups and institutes and offers some free software for sequence analysis. In the exercise section below, we will identify other software packages and their applications. A comprehensive compilation of bioinformatic applications available on the Web is published once a year in the journal *Nucleic Acids Research* [nar] (Web server issue).

3.4 Exercises

Exercise 3.1

Calculate the optimal alignment for the following sequences (◘ Fig. 3.3):
Sequence 1: MTPARGSALS
Sequence 2: MTPVRRSLS
Use the EMBOSS application Needle (► https://www.ebi.ac.uk/Tools/psa/emboss_needle/) to do this. Calculate the scores for the similarity matrices BLOSUM62, PAM250, and PAM30 using a gap open penalty and a gap extend penalty of 1. Do the suggested similarity matrices lead to similar alignments, or are there differences?

Exercise 3.2

Look for the Swiss-Prot database record for the human 5-hydroxytryptamine 2A receptor in the NCBI protein database [ncbi], and save the protein sequence in FASTA format.

Exercise 3.3

With the saved sequence from ► Exercise 3.2, perform a BLAST search for similar sequences in the nonredundant protein database of NCBI. Do this by going to the NCBI-BLAST page [ncbi-blast]. How many similar sequences are found? What information can be extracted from the graph on the results page?

Exercise 3.4

At the NCBI nucleotide database look for the entry with the AN AB037513 and save the nucleotide sequence in FASTA format. The sequence encodes a human 5HT2 receptor. Then perform BLAST searches using blastn and tblastx against the genome database of the organism *Drosophila melanogaster.* How many similar sequences are found in each case? What can be stated regarding the quality of the hits? What are the differences between the two programs blastn and tblastx, and how do the respective search results originate?

Exercise 3.5

Perform a local alignment of the protein sequences gi|543727 and gi|10726392 using *Global Align* at *Specialized Searches* on the NCBI-BLAST page [blast]. The ANs can be entered directly after selecting *Protein,* so that no further database queries are required. The two sequences are the already mentioned human 5HT2 receptor and its ortholog in *Drosophila melanogaster*. How can the result be interpreted?

Exercise 3.6

Perform a multiple alignment with the protein sequences gi|543727, gi|7296517 and NP_649806 using Clustal Omega [clustalomega]. How can the result be interpreted? As a remark: You have to download the sequences from the NCBI database and store them in FASTA format. The input mask on the Clustal Omage page [clustalomega] can be used unchanged with standard settings. The sequences can be inserted into the text mask.

Exercise 3.7

Perform a multiple alignment with the following sequences analogously to ► Exercise 3.6 and calculate a phylogenetic tree for the proteins Q28944.1, P25975.3, NP_081182.2, NP_640355.1, NP_001903.1, AAH12612.1. How can the result be interpreted? To what kind of proteins do the sequences belong? Note: Save both the alignment (*Download Alignment File*) and phylogenetic tree (*Download Phylogenetic Tree Data*). Have a look at the alignment with a visualizer. One visualizer can be found on the Clustal Omega page (tab *Phylogenetic Tree*). Treeview [treeview] is another online visualizer. You can directly copy the results (*Phylogenetic Tree* tab at *Tree Data*) into the first text field *Paste your tree in newick format*. Delete the second text field and display the tree using *ViewTree!*.

Exercise 3.8

Find an entry for a eukaryotic cosmid in the NCBI nucleotide database, e.g., AN: AC012088, and display the sequence in FASTA format. In a second browser window go to the Genscan server [genscan] and copy-paste the sequence into the corresponding window. Then run Genscan. Try to interpret the result. Search for further cosmid sequences of different species and repeat the exercise.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Boratyn GM, Camacho C, Cooper PS et al (2013) BLAST: a more efficient report with usability improvements. Nucleic Acids Res 41:W29–W33

Dayhoff MO, Schwartz RM, Orcutt BC (1978) In: Dayhoff MO (ed) Atlas of protein sequence and structure, Vol. 5, Suppl. 3. NBRF, Washington, DC, p 345

Eddy SR (2004) What is a hidden Markov model? Nat Biotechnol 10:1315–1316

Gerlt J, Babbitt P (2001) Respond: Orthologs and paralogs – we need to get it right. Genome Biol 2(8):1002.1–1002.3

Henikoff SB, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915–10919

Jensen RA (2001) Orthologs and paralogs – we need to get it right. Genome Biol 2(8):INTERACTIONS1002

Marchler-Bauer A, Derbyshire MK, Gonzales NR et al (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 4:2444–2448

Ramsden J (2015) Bioinformatics: An Introduction. Springer, New York City

Rice P, Longden I, Bleasby A (2000) EMBOSS: The european molecular biology open software suite. Trends Genet 16:276–277

Sievers F, Wilm A, Dineen D et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195–197

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 287: 631–637

Tatusova TA, Madden TL (1999) Blast 2 sequences – a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174:247–250

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF (1998) Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 26:3986–3990

Further Reading

bioedit. http://www.mbio.ncsu.edu/bioedit/bioedit.html
blast. https://blast.ncbi.nlm.nih.gov
clustalomega. http://www.ebi.ac.uk/Tools/msa/clustalo/
ddbj-blast. http://ddbj.nig.ac.jp/blast/blastn?lang=en
embnet. http://www.embnet.org/
embl-blast. https://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html
emboss. http://emboss.sourceforge.net/
expasy. https://www.expasy.org/
genscan. http://genes.mit.edu/GENSCAN.html
glimmer. http://ccb.jhu.edu/software/glimmer/index.shtml
ncbi. http://www.ncbi.nlm.nih.gov/
ncbi-blast. http://www.ncbi.nlm.nih.gov/blast/
sanger. http://www.sanger.ac.uk/
seaview. http://doua.prabi.fr/software/seaview
treeview. http://etetoolkit.org/treeview/