Bioinformatics: tools for analysis of biological sequences

Jan Pačes

Institute of Molecular Genetics, AS CR
Flemingovo 2, Prague
CZ-16637, Czech Republic

e-mail: hpaces@img.cas.cz

abbreviations: DNA - deoxyribonucleic acid; bp - base pair; contig - a long region of DNA assembled from shorter DNA sequences

1 Genomics and Bioinformatics

Information that directs all processes in living cells is stored in sequences of nucleotides in deoxyribonucleic acid (DNA). Contemporary methods of determining the nucleotide sequences, the so-called sequencing of DNA, are so effective that sequencing of whole genomes became feasible.

The field of Genomics is aimed at complex analysis of genomes based on our knowledge of the nucleotide sequences in DNA. Complete structures of several tenths of genomes have been determined so far (see http://kegg.genome.ad.jp/kegg/catalog/org_list.html or http://www.tigr.org/tdb/mdb). Most of them are bacterial genomes. These genomes usually consist from one chromosome sometimes accompanied by one or several plasmids. Bacterial chromosomes and plasmids are circular molecules of DNA. The number of nucleotides in prokaryotic (e.g. bacterial) genomes ranges from a fraction of million to several millions. Several genomes of higher eukaryotic organisms were also sequenced. They are the genomes of the yeast *Saccharomyces cerevisiae* (12 Mbp), the worm *Caenorhabditis elegans* (97 Mbp), the fruit fly *Drosophila melanogaster* (137 Mbp) and the plant *Arabidopsis thaliana* (116 Mbp). In 2001 nearly complete nucleotide sequence of the human DNA was announced. The complete human genome consists of more than three billion nucleotides. Several genome projects are completed every months. Most of them are bacterial genome projects, but genomes of several higher organisms (e.g. mouse or chimp) are getting close to completion.

Accumulation of the vast number of nucleotide sequences calls for a robust computer analysis. Bioinformatics is a new field devoted to analysis of long strings of nucleotide sequences generated in genome projects. Analysis of amino acid sequences of proteins encoded in the genomes usually follows the analysis of DNA and is an important part of bioinformatics.

To obtain the complete genome sequence it is necessary to assemble stepwise long nucleotide strings from the shorter nucleotide sequences usually generated from individual clones. In a typical example a sequence of several hundred nucleotides is determined in one sequence run. From these partial sequences longer and longer

strings (the so-called contigs) are assembled until the complete genome sequence is obtained.

The long contigs, in the ideal case the complete genome, are subjected to further computer analysis. We try to identify all genes present in the nucleotide sequence, to elucidate their structure (e.g. exon-intron organization), find elements regulating gene expression (e.g. promotors, enhancers, transcription terminators) and to identify other important DNA features. Genes are translated into the sequence of amino acids of the corresponding proteins. From these amino acid sequences basic features of the proteins are derived. This may for instance be the protein's secondary structure. Usually the overall DNA characteristics such as its base composition is also described.

After this basic DNA characterization nucleotide and amino acid sequences are often compared with the entries in international databases. Contemporary databases are very large. For instance the EMBL database of nucleotide sequences contains more then ten billion nucleotides of many genes and genomes. This number grows exponentially.

From the similarities found by this search we can often ascribe functions to individual genes and the corresponding proteins. The ultimate goal is to describe the complete metabolism of the organism. An important result of these comparisons are evolutionary relationships among organisms. It is now possible to describe on the molecule level individual taxons.

2 Features of Biological Sequences

Nucleotide and amino acid sequences have several special features that have to be taken into account when performing computer analysis. These special features are connected with the biochemical and biological function of genes and proteins. For instance, variations in nucleotide sequences performing the same function (e.g. promoters) make the analysis difficult.

2.1 DNA

DNA is the polymer molecule in which genetic information of organisms is stored. DNA consists of four basic components, the so-called nucleotides. Each nucleotide consists of a sugar deoxyribose, a residue of phosphate and one of the four nitrogenous bases. These basis are adenine (A), guanine (G), cytosine (C) and thymine (T). Nucleotides are connected by sugar-phosphate bonds into long strings. DNA is composed from two strings that run antiparallel in the well known double helix. The two strings (strands) are complementary. This means that A in one strand pairs through hydrogen bonds with T in the other strand and G pairs with C. Genetic information stored in the sequence of A, C, G and T reads in one direction only. Because the two strands are antiparallel genetic information reads in the two strands in opposite direction. For instance the sequence ATTGCA in one strand reads TGCAAT in the complementary strand.

In databases and for computer analysis the DNA sequences are stored in the single-letter code. Because of occasional ambiguities in sequence analysis additional letters are used to facilitate nucleotide analysis (Tab. 1).

Table 1: Nucleotide code. Compl. stands for complementary nucleotide.

name	code	nucleotide	compl.
Adenine	A	A	T
Cytosine	C	C	G
Guanine	G	G	C
Thymine	T	T	A
Uracil	U	U	A

code	nucleotide	compl.
M	A/C	K
R	A/G	Y
W	A/T	S
S	C/G	W
Y	C/T	R
K	G/T	M
V	A/C/G	B
H	A/C/T	D
D	A/G/T	H
B	C/G/T	V
N	A/C/G/T	N
-	space	-

2.2 Translation of DNA to Proteins

Regions of DNA to which most attention is devoted are genes. In prokaryotic cells genes usually consist of uninterrupted nucleotide sequence. In contrast, genetic information of eukaryotic organisms is organized in a more complex fashion. Regions encoding amino acids (exons) are interspaced by long non-coding sequences (introns). Thus only two to three percent of human DNA encodes proteins.

A sequence of three nucleotides, the so-called triplet or codon, determine which amino acid is incorporated in the corresponding protein. Proteins consist of 20 types of two amino acids. With the exception of two amino acids all are encoded by more then one codon. This implies that although translation of DNA to protein is unique it is impossible to unequivocally derive nucleotide sequence from the amino acid sequence. In addition, a region of DNA can encode six different proteins.

Genetic code is in Table 2.

2.3 Proteins

Proteins are the functional molecules operating in cells. Basic building blocks of proteins are amino acids. Proteins consists of 20 types of amino acids. For amino acids a one letter code is used, although the older three letter code can be occasionally also found in biochemical literature (Tab. 3).

From the biochemical point of view it is important that unlike nucleotides amino acids generally are much more chemically different. However, some of them are so similar that they can replace each other in functional proteins. For example leucine, isoleucine and valine are often found in the same position in functionally identical proteins isolated from different organisms.

3 Main types of analyses

The most frequent tasks of DNA and protein analysis are:

Table 2: Genetic code

1	2	A	G	C	T		3
	A	Lys	Arg	Thr	Ile	A	
		Lys	Arg	Thr	Met	G	
		Asn	Ser	Thr	Ile	C	
		Asn	Ser	Thr	Ile	T	
	G	Glu	Gly	Ala	Val	A	
		Glu	Gly	Ala	Val	G	
		Asp	Gly	Ala	Val	C	
		Asp	Gly	Ala	Val	T	
	C	Gln	Arg	Pro	Leu	A	
		Gln	Arg	Pro	Leu	G	
		His	Arg	Pro	Leu	C	
		His	Arg	Pro	Leu	T	
	T	*	*	Ser	Leu	A	
		*	Trp	Ser	Leu	G	
		Tyr	Cys	Ser	Phe	C	
		Tyr	Cys	Ser	Phe	T	

Table 3: Amino acid code

1-code	3-code	amino acid
A	Ala	alanine
C	Cys	cysteine
D	Asp	asparagic acid
E	Glu	glutamic acid
F	Phe	phenylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine

1-code	3-code	amino acid
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	serine
T	Thr	threonine
V	Val	valine
W	Trp	tryptophan
Y	Tyr	tyrosine
B	Asx	aspartic acid or asparagine
Z	Glx	glutamic acid or glutamine
X	Xxx	any amino acid
*	—	stop

- Assembly: to assemble contigs from short (several hundred nucleotide long) sequences.
- Gene prediction: to identify genes in DNA.
- Pattern search: to find regions composed of typical short sequences.
- Pairwise alignment: to find in a database similar or homologous sequence.
- Multiple alignment: to assess relationship of several sequences.

4 Comparison of Biological Sequences

The basic step in analyzing a determined nucleotide sequence is its comparison with the sequences already deposited in databases. We are looking for related and/or similar sequences. The presumption for this search is that in evolution nucleotide substitutions, and less frequently nucleotide deletions and insertions, are accumulated. Thus similarity of genes and proteins can be traced in more or less distantly related organisms. Accumulation of these changes is not random. It is more frequent in the DNA regions encoding those part of proteins that are not fundamentally important for the protein's function.

An algorithm for pairwise comparison of amino acid sequences was described for the first time by Needleman & Wunsch in [1] and is known as the global Needleman-Wunsch algorithm. Because in most cases two sequences are more similar in certain regions and less similar in other in other regions, the Needleman-Wunsch algorithm was improved for evaluation of local similarities. This is called local Smith-Waterman alignment [2] [3] [4].

The basic principle of the similarity search is known as the pairwise alignment and is based on comparison of sequences pair by pair and on search for the best alignment with highest similarity. Score of bonuses and penalizations for matches, mismatches, deletions etc. are calculated for all possible pairs. Scores of all alignments are then calculated as the sum of scores of all pairs in the alignment. The best similarity it then ascribed to the alignment with the highest score. This is known as the Smith-Waterman score and is usually used as basic characteristics of the alignment.

4.1 Scoring Matrix

When comparing two nucleotide sequences, it is not necessary to evaluate similarity of individual nucleotides. It is sufficient to consider identities. Fundamentally different are comparisons of amino acid sequences. When analyzing evolutionarily related proteins it was discovered that for enzyme activities general biochemical properties of individual amino acids are very important. For many comparisons it is useful to group amino acids according to their chemical properties such as hydrophobicity, charge, size, polarity etc. Substitution of amino acids belonging to one such group may be penalized less compared to substitutions of unrelated amino acids.

However, it is also important to take into account genetic (evolutionary) relatedness of individual amino acids. For instance, tryptophane is encoded by the TGG codon. One mutation leads to codons for glycine (GGG), serine (TCG) and leucine (TTG), two codons for cysteine (TGT, TGC), arginine (CGG, AGG) and two stop codons (TGA, TAG). Conversion of tryptophane to arginine is therefore more likely than conversion to glycine in spite of the fact that tryptophane and arginine are chemically very different: tryptophane is hydrophobic aromatic amino acid and arginine is hydrophilic polar positively charged amino acid.

These considerations are taken into account in the so-called scoring matrix, which is basically evaluation of a replacement of one amino acid by another amino acid. Today we use two types of scoring matrix: PAM and BLOSUM. PAM and BLOSUM differ by the calculation method and they give similar results. Table 4 shows a part of the scoring matrix concerning tryptophane (W).

Table 4: Scoring matrix for tryptophane (W) and its change to selected amino acids (R,N,D,C).

W	R	N	D	C	W
PAM 50	-1	-7	-12	-13	13
PAM 100	1	-5	-9	-9	12
PAM 250	2	-4	-7	-8	17
BLOSUM 100	-7	-8	-10	-7	17
BLOSUM 62	-3	-4	-4	-2	11
BLOSUM 30	0	-7	-4	-2	20

5 Often Used Programs

Demands for memory and computational time grows with the length of the biological sequence under evaluation and with volume of the database. This is why an important principle is pre-selection of possible hits. This pre-selection is based on observation that two evolutionarily distant sequences usually contain conserved short regions. These conserved regions can served for fishing out evolutionarily related sequences. Most commonly used programs based on this principle are FASTA and BLAST, which are designed to compare both nucleotide and amino acid sequences. Moreover, these two programs deal with a basic feature of genetic information, i.e. conversion of the language of the nucleotide sequences in DNA into the language of amino acids sequences in proteins. This is done in the six possible reading frames.

5.1 FASTA

The FASTA program (FAST Algorithm) [5] was one of the first freely available programs of bioinformatics. FASTA first prepares list of "words", i.e. very short portions of the sequence in question. This list is then compared with the entries in the database. If several words match an entry in the database in the right order and close-by the sequence is selected for the complete Smith-Waterman alignment.

There are several variants of the FASTA program, according to the type of the sequence and database used. The main FASTA program is used for searching nucleotide sequences in DNA databases or amino acid sequences in protein databases. FASTX/FASTY are used for comparison od DNA against proteins. TFASTX/TFASTY can be used to look for proteins in DNA databases.

An important parameter of the search is "expectancy" (E). E is proportional to the probability with which the same degree of similarity can be found in a random sequence of the same length. Because Smith-Waterman score depends on the sequence length it was normalized to the unit of length (so-called "bit score")

Next is an example of the FASTA search.

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t09 May 18, 2001

428405286 residues in 67627 sequences
 statistics extrapolated from 60000 to 69642 sequences
  Expectation_n fit: rho(ln(x))= 13.2503+/-0.000153; mu= -20.9826+/- 0.010
 mean_var=452.3589+/-70.442, 0's: 3 Z-trim: 470  B-trim: 62 in 1/86
```

```
 Lambda= 0.0603

FASTA (3.39 May 2001) function [optimized, +5/-4 matrix (5:-4)] ktup: 6
 join: 73, opt: 58, gap-pen: -16/ -4, width:  16
 Scan time: 16.480
The best scores are:                                      opt bits E(69642)
EM_INV:DM19269 U19269 Drosophila melanogaste (4976)  [f] 736   80 6.2e-13
EM_INV:AC007185 AC007185 Drosophila melanoga (77732) [r] 723   80 7.8e-13
EM_INV:AF139019 AF139019 Cepaea nemoralis mi ( 624)  [r] 452   55 2.6e-05
...

>>EM_INV:DM19269 U19269 Drosophila melanogaster Dachshun  (4976 nt)
 initn: 870 init1: 692 opt: 736  Z-score: 352.9  bits: 80.3 E(): 6.2e-13
 59.627% identity (62.036% ungapped) in 644 nt overlap (242-877:976-1602)

             220       230       240       250       260       270
gi     CAGTCACCTCTCCTGGTGGCGGCGGCGGCAGCGGAGGCGGCGGTGGCAGCGGCGGCA
                                    ::::: :: :: ::     :::: ::  ::
EM_INV TCCGGTGAGCTCCCTCAACCACTCCATGATGCAGCAGATGCAGC---AACAGCAGCAACA
         950       960       970       980           990       1000

             280       290       300       310       320       330
gi     ACGGAGGCGGCGGCGGGAGCAACTGCAACCCCAGCCTGGCGGCCGGGAGCAGCGGCGGCG
       :: :  :: ::  : : :::: : ::::: ::: :    : :::    ::  :   :
EM_INV ACAGCAGCAGCAACAGCAGCAGCAGCAACACCATCAGCTCAGCCCCCCGCCACATGGAAT
           1010      1020      1030      1040      1050      1060

             340          350       360       370       380
gi     GCGTTAGCG---CTGGCGGCGGCGGCGCCTCCAGCACCCCCATCACCGCGAGCACCGGCA
       ::  : : :   : : : :: :  : :::: : :: ::   ::  ::     ::  : :
EM_INV GCCATCGGGCAACGGACTGCCGACGGGCCTACCGC-CCAGAATGCCC-----CATGGACT
           1070      1080      1090       1100           1110

...
```

5.2 BLAST

BLAST (Basic Local Alignment Search Tool) [6] [7] [8] was developed in 1990. BLAST first maps the database for presence of various strings and lists them. Then, similarly to FASTA, it creates a list of "words" from the searched sequence. These words are then compared to the list of strings. This leads to selection of the best hits which are then extended from both sides. The BLAST program has several variants similarly to FASTA.

6 Multiple Alignment

A very important task in evaluating biological sequences is comparison of more then two sequences at the same time. This is called multiple alignment. Multiple alignment helps to identify biologically important parts of genes. In addition it enables to estimate evolutionary distances among biological sequences, to formulate consensus sequences and parental sequences. Especially important are consensus sequences because they can serve in identification of additional, often more distant members

of the gene family in question.

The most common program for the multiple alignments is CLUSTALW ([9] [10] [11] [12]). CLUSTALW compares by the Needleman-Wunsch algorithm all sequences in the query and it selects the most similar pair. From these two sequences a consensus sequence is generated and it is used for aligning another sequence. With this stepwise mechanism whole family of sequences are compared. The relatedness of individual members of a sequence family can be assessed..

An example of the CLUSTALW search follows.

```
Hs-U3   LDALSRECCVTAGGRDGTVRVWKI----PEESQLVFYGH--------------------------
Mm-U3   LDALSRECCVTAGGRDGTVRVWKI----PEESQLVFYGH--------------------------
Xl-U3   LDSLSRERCVTVGGRDGTMRIWKI----AEETQLVFSGH--------------------------
At-U3a  IDALRKERALTVG-RDRTMLYHKV----PESTRMIYRAP--------------------------
At-U3b  IDALGRERVLSVG-RDRTMQLYKVGIVVPESTRLIYRAS--------------------------
Dm-U3   IDALSRERAITAGGSDCSLRIWKI----TEESQLIYNGH--------------------------
Sp-U3   VDALARERCVSVGGRDRTSRLWKI----VEESQLVFRSGGTSMKAT-----AGYM-----------
Nc-U3   IDALAGERCVSVGARDRTARYWKV----PEESQLVFRGGVSEKKSHKNRDQAVNH-----------
Ce-U3   IGVLSKQRVATVGGRDRSARLWKV----EDESQLMFSGLQN------------------------
Sc-U3   ISALAMERCVTVGARDRTAMLWKI----PDETRLTFRGGDEPQKLLRRWMKENAKEGEDGEVKYPD
        :. *  :   :.*  * :    *:     :.::: : .

Hs-U3   ---------QGSIDCIHLINEEHMVSGADDGSVALWGLSKKRPLALQREAHGLRGE----------
Mm-U3   ---------QGSIDCIHLINEEHMVSGADDGSVALWGLSKKRPLALQREAHGLHGE----------
Xl-U3   ---------EGSIDCVRLINEEHIVTGADDGSLALWTVGKKKPLTQMKCAHGSYGE----------
At-U3a  ---------ASSLESCCFISDNEYLSGSDNGTVALWGMLKKKPVFVFKNAHQDIPDGITTNGILEN
At-U3b  ---------ESNFECCCFVNSDEFLSGSDNGSIALWSILKKKPVFIVNNAHHVIAD----------
Dm-U3   ---------KDSIECVKYINDEHFVSGGMDGAIGLWSALKKKPICTTQLAHGVGEN----------
Sp-U3   ---------EGSVDCVAMIDEDHFVTGSDNGVIALWSVQRKKPLFTYPLAHGLDPILAPGRHSAET
Nc-U3   ---------DGTMDQVAMIDDELFVTGDSAGTLSLWGINRKKALFTQPCAHGIDPPLKPTEVSADA
Ce-U3   ---------CVSLDCVAMINEEHFATGSADGSIALWSFWKKRALHVRKQAHGTQNG----------
Sc-U3   ESEAPLFFCEGSIDVVSMVDDFHFITGSDNGNICLWSLAKKKPIFTERIAHGILPEPSFNDISGET
                  ..:    :..    :*   * : **   :*:.:     **
```

References

[1] Wunsch CD Needleman SB. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3), 1970.

[2] Waterman MS Smith TF. Identification of common molecular subsequences. *J Mol Biol*, 147(1), 1981.

[3] Taylor P. A fast homology program for aligning biological sequences. *Nucleic Acids Res*, 12(1 Pt 2), 1984.

[4] Waterman MS. Efficient sequence alignment algorithms. *J Theor Biol*, 108(3), 1984.

[5] Lipman DJ Pearson WR. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8), 1988.

[6] Altschul SF et al. Basic local alignment search tool. *J Mol Biol*, 215(3), 1990.

[7] States DJ Gish W. Identification of protein coding regions by database similarity search. *Nat Genet*, 3(3), 1993.

[8] Altschul SF et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17), 1997.

[9] Sharp PM Higgins DG. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1), 1989.

[10] Fuchs R Higgins DG, Bleasby AJ. Clustal v: improved software for multiple sequence alignment. *Comput Appl Biosci*, 8(2), 1992.

[11] Higgins DG. Clustal v: multiple alignment of dna and protein sequences. *Methods Mol Biol*, 25, 1994.

[12] Gibson TJ Thompson JD, Higgins DG. 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(22), 1995.

Acknowledgements: This work was supported by Center of Integrated Genomics and by grant GA301/99/M023 of the Grant Agency of the Czech Republic.