Bioinformatics

An outline

Krishnendu Sinha
Assistant Professor of Zoology
Jhargram Raj College

Outline of the discussion

1. INTRODUCTION
2. INFORMATION NETWORK
3. PROTEIN INFORMATION RESOURCE
4. GENOME INFORMATION RESOURCE
5. PAIRWISE ALIGNMENT TECHNIQUES

1. INTRODUCTION

- The application of computational techniques in the management and analysis of biological information is known as *Bioinformatics*
- Rybak in 1968 coined the term in his book
- The sequence databases is doubling, approximately each year
- The key challenge is to manage and analyse these immense overwhelming information to draw a logical conclusion in respect to protein structure, function and evolution *(also true for nucleotide information!!!)*
- Two principle analytical approach in bioinformatics are,
  i) pattern-recognition & ii) prediction
- The barrier in prediction is the **Protein Folding Problem**

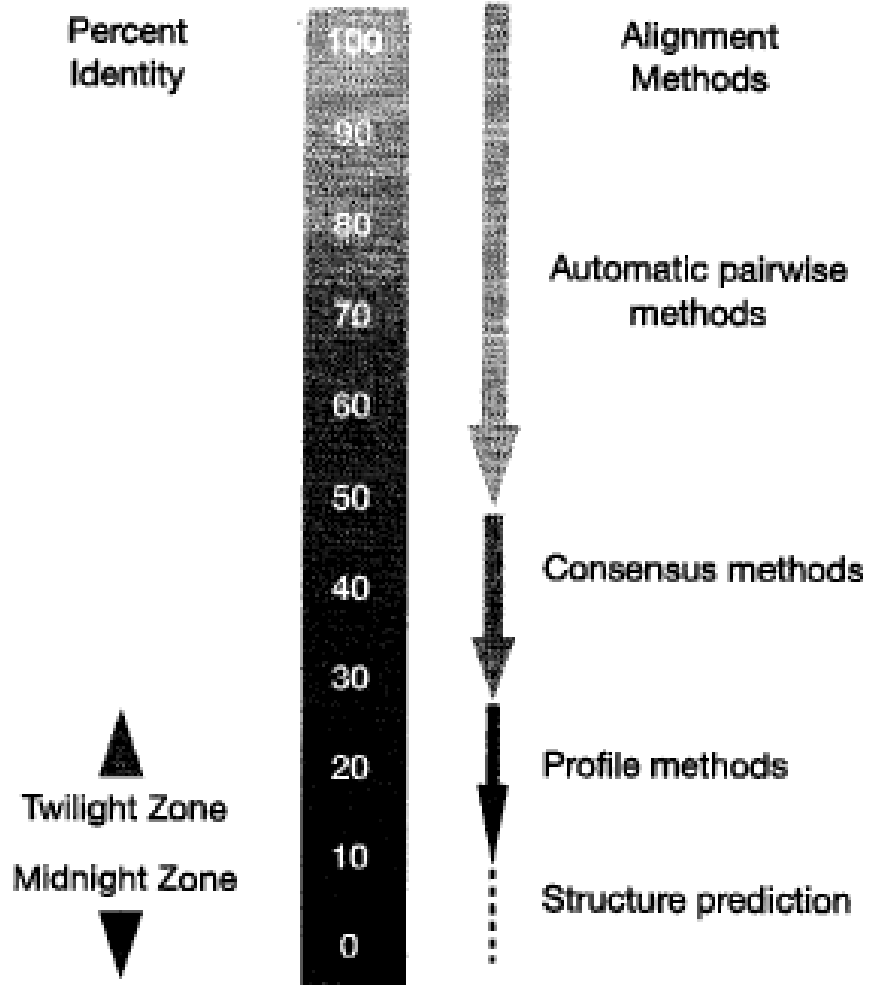
- **Homology** is the central concept used in sequence analysis *(homology is not synonymous to similarity)*

- Sequences are said to be homologous if they are said to be diverged from a common ancestor

- Homology could be of two types, i) **paralogous** (different but related functions in same species), ii) **orthologous** (same function in different species)

- **Analogy** is another important concept developed from convergent evolution (non-homologous proteins having similar functions/ similar protein folds with no detectable sequence similarity)

- Alignment searches can be with decreasing certainty towards the **Twilight Zone** *(the zone of sequence similarity [0-20% aprox]where alignment appears to be plausible in naked eye)* where alignments are no longer statistically significant where as comparison fails completely in the **Midnight Zone**



Application areas of different analysis methods. The scale indicates percent identity between two aligned sequences. Alignment of random sequences can produce around 20% identity; less than 20% does not constitute a significant alignment. Around this threshold is the Twilight zone, where alignments may appear plausible to the eye, but cannot be proved by current methods. Beyond the Twilight Zone is the so-called Midnight Zone, where sequence comparisons fail completely to detect structural similarities.

2. INFORMATION NETWORK

Browser: A Web client (computer program) that permits information retrival from the Internet or the WWW

Client: Any program that interacts with a server (e.g. firefox)

Server: A computer or software system that communicates information via Internet to a client

Transmission Control Protocol/Internet Protocol (TCP/IP): The rules that govern data transmission between two computers over the Internet

Internet Protocol address (IP address): An unique identifying number assigned to each node on internet to allow communication between them

HyperText Markup Language (HTML): The syntax governing the way documents are created so that they can be interpreted and rendered by Web browsers

HyperText Transport Protocol (HTTP): The communication protocols used by Web servers

Hypertext: Text that contains embedded links hyperlinks to other documents

Hyperlink: An active HTTP cross-reference that link one Web document to another on the Internet

- **Internet** is a global network of computer networks
- Each computer in the network is called the node and each **node** has an unique **id (IP address)** by which it can be identified and can communicate with other such node
- Internet provide services like emails, news groups, file transfer, remote computing etc
- The **World Wide Web (WWW)** is the most powerful information system on the internet *(but its not the same as internet!!!)*
- **Browsers** provide easy-to-use interface for accessing information on the Web
- Home page is the first point of contact between a browser and a Web server
- Documents that browsers displayed are accessed by means of unique address called URLs (Uniform Resource Locators).

Example Internet domains and subdomains

Country-based domains		*Other domains*		*Subdomains*	
Australia	.au	Educational	.edu	Academic	.ac
Denmark	.dk	Commercial	.com	Company	.co
Finland	.fi	Governmental	.gov	Other organisation	.org
France	.fr	Military	.mil	General	.gen
Germany	.de				
Greece	.gr				
Hungary	.hu				
Ireland	.ie				
Israel	.il				
Italy	.it				
Netherlands	.nl				
New Zealand	.nz				
Poland	.pl				
Portugal	.pt				
South Africa	.za				
Spain	.es				
Sweden	.se				
Switzerland	.ch				
United Kingdom	.uk				
USA	.us				

European Molecular Biology Network (EMBnet)

- EMBnet is a network of European Biocomputing laboratories established in 1988

- Structurally it is subdivided into cluster of nodes called **National Nodes** *(online service, user support and training)*, **Specialist Nodes** *(database management and software development)* **and Associate Node**

- Such three very notable Specialist Nodes are hosted by **Hinxton Hall at Wellcome Trust Genome Campus**

- These are the **Sanger Centre, the UK MRC Human Genome Mapping ProjectResource Centre (HGMP-RC) and the European Bioinformatics Institute** *(an outstation of EMBL maintain, EMBL nucleotide database, TrEMBL and SWISS-PROT database; also collaborate with GenBank and DDBJ as a member of a common collaborative work)*

- The **Sequence Retrieval System (SRS)** was developed within EMBnet to allow information retrieval across a range of different database types by using a single interface

EMBnet National Nodes

Vienna Biocenter	Austria	http://www.at.embnet.org/
BEN	Belgium	http://www.be.embnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.embnet.org/
INFOBIOGEN	France	http://www.infobiogen.fr/
GENIUSnet	Germany	http://genome.dkfz-heidelberg.de/biounit/
IMBB	Greece	http://www.imbb.forth.gr/
HEN	Hungary	http://www.hu.embnet.org/
INCBI	Ireland	http://acer.gen.tcd.ie/
INN	Israel	http://dapsas.weizmann.ac.il/bcd/inn.html
IEN-ADR	Italy	http://bio-www.ba.cnr.it:8000/BioWWW/Bio-WWW.htm
CAOS/CAMM	Netherlands	http://www.caos.kun.nl/
Bio	Norway	http://www.no.embnet.org/
IBB	Poland	http://www.ibb.waw.pl/
IGC	Portugal	http://www.igc.gulbenkian.pt/
GeneBee	Russia	http://www.genebee.msu.su/
CNB-CSIC	Spain	http://www.es.embnet.org/
BMC	Sweden	http://www.embnet.se/
SIB	Switzerland	http://www.ch.embnet.org/
SEQNET	UK	http://www.seqnet.dl.ac.uk/

EMBnet Specialist Nodes

MIPS	Germany	http://www.mips.biochem.mpg.de/
ICGEB	Italy	http://www.icgeb.trieste.it/
Pharmacia Upjohn	Sweden	http://www.pnu.com/
F.Hoffmann-La Roche	Switzerland	http://www.roche.com/
EBI	UK	http://www.ebi.ac.uk/
HGMP-RC	UK	http://www.hgmp.mrc.ac.uk/
Sanger	UK	http://www.sanger.ac.uk/
UMBER	UK	http://www.bioinf.man.ac.uk/dbbrowser

EMBnet Associate Nodes

IBBM	Argentina	http://sol.biol.unlp.edu.ar/embnet
ANGIS	Australia	http://www.angis.su.oz.au/
CBI	China	http://www.cbi.pku.edu.cn/
CIGB	Cuba	http://bio.cigb.edu.cu/
CDFD	India	http://salarjung.embnet.org.in/
SANBI	South Africa	http://www.sanbi.ac.za

National Centre for Biotechnology Information (NCBI)

- Leading American bio-information provider and home of the **GenBank** and **Entrez** information retrieval system
- Established in the year of 1988, hosted by **National Library of Medicine (NLM)** and situated in the campus of **National Institute of Health (NIH)**, Bethesda, Maryland
- The main function is to maintain the GenBank and NIH DNA sequence database and also collaborating with EMBL and DDBJ

USA Information Providers

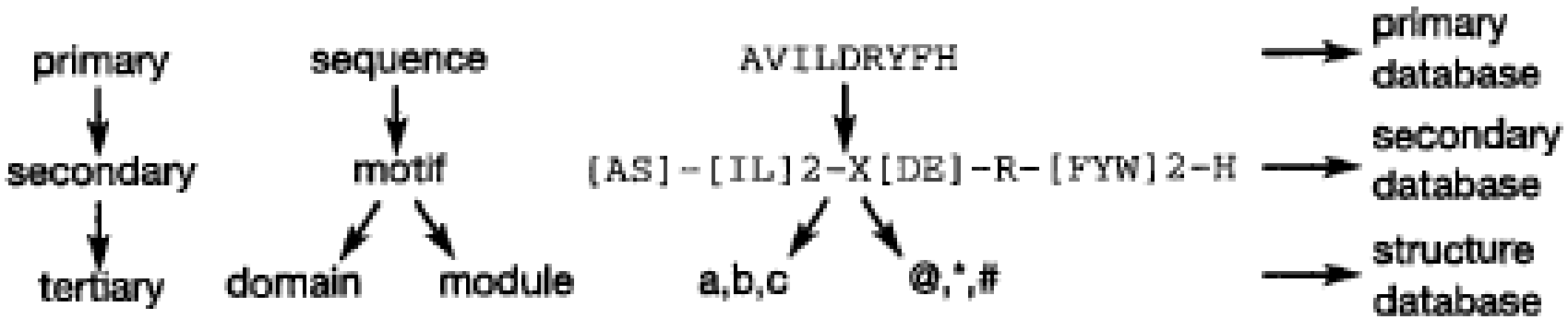
NCBI	USA	http://www.ncbi.nlm.nih.gov/
NLM	USA	http://www.nlm.nih.gov/
NIH	USA	http://www.nih.gov/

Biotechnology Information System Network (BTISNet)

The Indian bioinformatics network

3. PROTEIN INFORMATION RESOURCE

- Databases are used to store a vast amount of data generating from different sequence projects
- Among different types of databases **primary, secondary and tertiary databases** are of most importance for routine sequence analysis
- Primary databases contains **sequence data**
- Composite databases **amalgamate different primary databases** and thus obviate the need of searching different databases for single query
- Different composite databases use different primary resources and have different redundancy criteria for its amalgamation process
- Secondary databases contains **pattern data** (diagnostic signatures of protein families). These signatures encode the most highly conserved features of multiple alignment data and often crucial for structure and function analysis of proteins
- Different sequence analysis methods gives rise to different pattern databases: the main approaches exploit single motifs (regular expressions), multiple motif (e.g. fingerprints) and full domain alignment (e.g. Hidden Markov Models)
- **PROSITE** and **PRINTS** are the only and comprehensively manually annotated secondary databases
- Unified database for protein family is known as **InterPro,** created to avoid annotation bottleneck of the secondary databases



Primary nucleic acid and protein sequence databases.

Nucleic acid	*Protein*
EMBL	PIR
GenBank	MIPS
DDBJ	SWISS-PROT
	TrEMBL
	NRL-3D

Some of the available composite protein sequence databases, with details of their primary data sources.

NRDB	*OWL*	*MIPSX*	*SP+TrEMBL*
PDB	SWISS-PROT	PIR1–4	SWISS-PROT
SWISS-PROT	PIR	MIPSOwn	TrEMBL
PIR	GenBank	MIPSTrn	
GenPept	NRL-3D	MIPSH	
SWISS-PROTupdate		PIRMOD	
GenPeptupdate		NRL-3D	
		SWISS-PROT	
		EMTrans	
		GBTrans	
		Kabat	
		PseqIP	

Some of the major secondary 'pattern' databases: in each case, the primary source is noted, together with the type of pattern stored. PRINTS is currently the only secondary resource to be derived from a composite.

Secondary database	*Primary source*	*Stored information*
PROSITE	SWISS-PROT	Regular expressions (patterns)
Profiles	SWISS-PROT	Weighted matrices (profiles)
PRINTS	OWL*	Aligned motifs (fingerprints)
Pfam	SWISS-PROT	Hidden Markov Models (HMMs)
BLOCKS	PROSITE/PRINTS	Aligned motifs (blocks)
IDENTIFY	BLOCKS/PRINTS	Fuzzy regular expressions (patterns)

*SWISS-PROT is OWL's highest priority source.

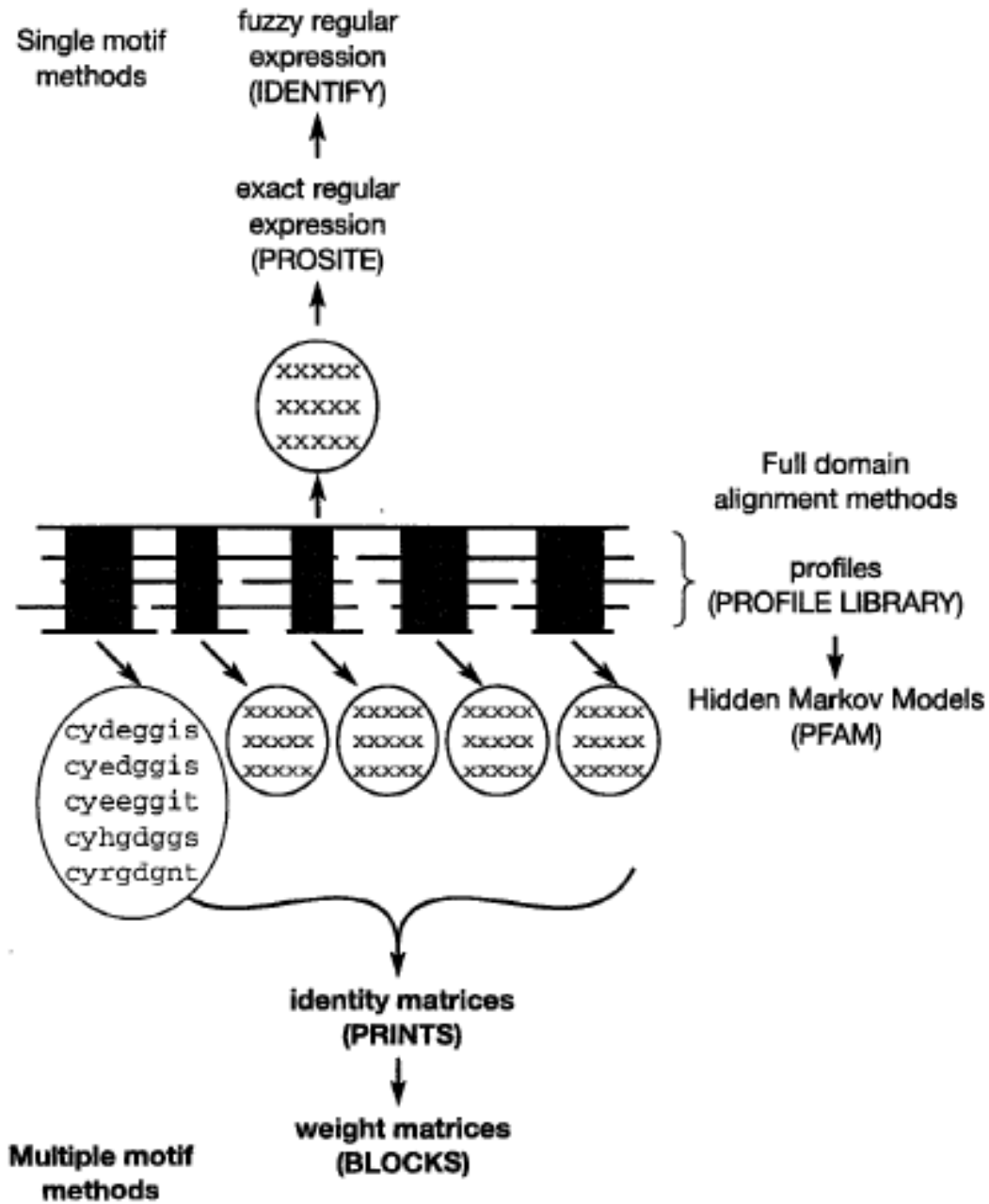


Illustration of the three principal methods for building pattern databases, i.e., using single motifs, multiple motifs and full domain alignments.

4. GENOME INFORMATION RESOURCE

```
LOCUS       HUMCYCLOX  3387 bp  mRNA  PRI  31-DEC-1994
DEFINITION  Homo sapiens cyclooxygenase-2 (Cox-2) mRNA, complete cds.
ACCESSION   M90100
NID         g181253
KEYWORDS    cyclooxygenase-2; prostaglandin synthase.
SOURCE      Homo sapiens umbilical vein cDNA to mRNA.
  ORGANISM  Homo sapiens
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1 (bases 1 to 3387)
  AUTHORS   Hla,T. and Neilson,K.
  TITLE     Human cyclooxygenase-2 cDNA
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 89 (16), 7384-7388 (1992)
  MEDLINE   92366465
FEATURES             Location/Qualifiers
     source          1..3387
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /cell_type="endothelial"
                     /tissue_type="umbilical vein"
       5'UTR         1..97
                     /gene="Cox-2"
       gene          1..3387
                     /gene="Cox-2"
       CDS           98..1912
                     /gene="Cox-2"
                     /EC_number="1.14.99.1"
                     /codon_start=1
                     /product="cyclooxygenase-2"
                     /db_xref="PID:g181254"
                        /translation="MLARALLLCAVLALSHTANPCCSHPCQNRGVCMSVGFDQYKCDC
                        TRTGFYGENCSTPEFLTRIKLFLKPTPNTVHYILTHFKGFWNVVNNIPFLRNAIMSYV
                        ...
                        EYRKRFMLKPYESFEELTGEKEMSAELEALYGDIDAVELYPALLVEKPRPDAIFGETM
                        VEVGAPFSLKGLMGNVICSPAYWKPSTFGGEVGFQIINTASIQSLICNNVKGCPFTSF
                       SVPDPELIKTVTINASSSRSGLDDINPTVLLKERSTEL"
       sig_peptide   98..148
                     /gene="Cox-2"
       mat_peptide   149..1909
                     /gene="Cox-2"
                     /EC_number="1.14.99.1"
                     /product="cyclooxygenase-2"
       3'UTR         1913..3387
                     /gene="Cox-2"
       polyA_signal  3369..3374
                     /gene="Cox-2"
BASE COUNT       1010 a    712 c    633 g    1032 t
ORIGIN
   1 gtccaggaac tcctcagcag cgcctccttc agctccacag ccagacgccc tcagacagca
  61 aagcctaccc cgcgccgcg cctgcccgc cgctgcgatg ctcgcccgcg ccctgctgct
 121 gtgcgcggtc ctggcgctca gccatacagc aaatccttgc tgttcccacc catgtcaaaa
     ...
3301 tacctgaact tttgcaagtt ttcaggtaaa cctcagctca ggactgctat ttagctcctc
3361 ttaagaagat taaaaaaaaa aaaaaag
//
```

Example GenBank entry illustrating the use of keywords, sub-keywords and the Feature Table to express information on the structure of the cDNA for Cox-2. Both the protein translation in the Feature Table and the nucleotide sequence have been abbreviated (...) for the figure.

- The principle DNA sequence databases are **GenBank, EMBL and DDBJ**, which each collect a portion of the total sequence data reported world wide and exchange them at a daily basis

- GenBank is produced at NCBI and splitted into smaller discrete divisions. This allows fast, specific searches by restricting queries to particular database subsets

The three-letter codes for each of the 17 divisions of GenBank.

Division	*Sequence subset*
PRI	Primate
ROD	Rodent
MAM	Other mammalian
VRT	Other vertebrate
INV	Invertebrate
PLN	Plant, fungal, algal
BCT	Bacterial
RNA	Structural RNA
VRL	Viral
PHG	Bacteriophage
SYN	Synthetic
UNA	Unannotated
EST	EST (Expressed Sequence Tags)
PAT	Patent
STS	STS (Sequence Tagged Sites)
GSS	GSS (Genome Survey Sequences)
HTG	HTG (High Throughput Genomic Sequences)

- In addition to these comprehensive DNA sequence databases, there is a variety of more specialised specific genomic resources often termed boutique databases

- These bring focus to specific genomics and particular genomics techniques

- **SGD (*Saccaromyces* Genome Database)**

- **UniGene:** primarily attempts to provide a transcript map by utilizing set of non-redundant gene oriented clusters derived fro GeneBank sequences

- **TDB (TIGR database):** Microbes database

- **ACeBD (A *C. elegance* database)**

5. PAIRWISE ALIGNMENT TECHNIQUES

- To identify an evolutionary relationship between a newly determined sequence and a known gene family, the extent of shared similarity must be assessed

- An **algorithm** is a set of finite steps that define a computational process

- A **program** is an implementation of algorithm

- The simplest way to compare two sequences is to align them by inserting gap characters to bring them to vertical register. Counting the matched character positions gives a naïve alignment score

- The basic method of comparing two sequences are dotplot. This is a graph where sequence lye in x and y-axes.

- Dots are plotted in all positions where identical residues are observed. For identical sequences, this leads to an unbroken diagonal line across the plot, where similar sequences given rise to broken diagonals

- Alignments are models that reflects different biological perspectives. Therefore, there is no right or wrong model from one another. Two general approaches consider similarity (1) **global alignment** (across the full length of the sequences-the **Needleman and Wunsch algorithm**) and (2**) local alignment** (across only parts of the sequences-**Smith-Waterman algorithm**)

- Both the algorithm exploit dynamic programming, whereby a solution to a problem is built by solving smaller, tractable sub-problems. The optimal alignment is chosen from a set of high-scoring alternatives. Such methods are prohibitively time consuming for a larger pair of sequence.

- The FastA and BLAST programs are local similarity search methods that concentrate on finding short identical matches, which may contribute to a total match

- Speed issues are addressed using heuristics

```
Unaligned
Sequence 1 (query)    AGGVLIIQVG
                      ||||||
Sequence 2 (subject) AGGVLIQVG

Aligned
Sequence 1 (query)    AGGVLIIQVG
                      ||||| ||||
Sequence 2 (subject) AGGVL-IQVG
```

Illustration of the use of a gap character '-' to bring two sequences into alignment; vertical bars denote identical matches – six in the first alignment, nine in the second.

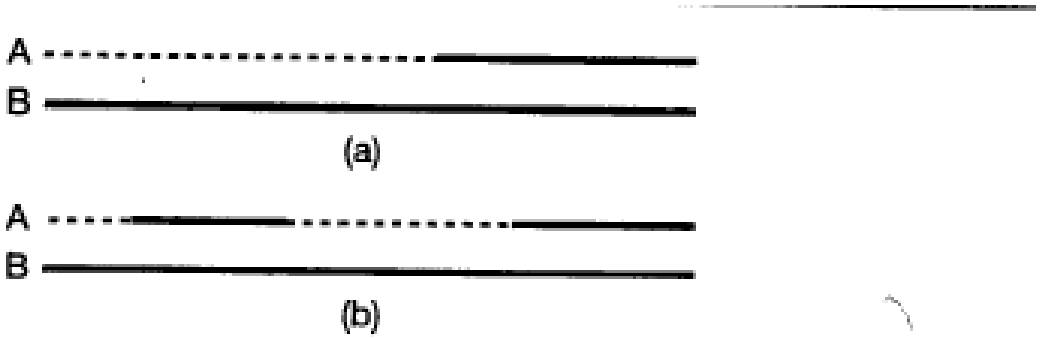


Illustration of the alignment of a sub-sequence A with a full-length sequence B, showing: (a) the situation where A is identical to one part of B, and insertion of one block of gaps allows complete alignment of the two sequences; and (b) the situation where A is identical to different parts of B, so that more than one block of gaps must be inserted to bring the sequences into register.

Unitary scoring matrices: (a) DNA and (b) protein – the amino acids are grouped according to their physicochemical properties.

(a)

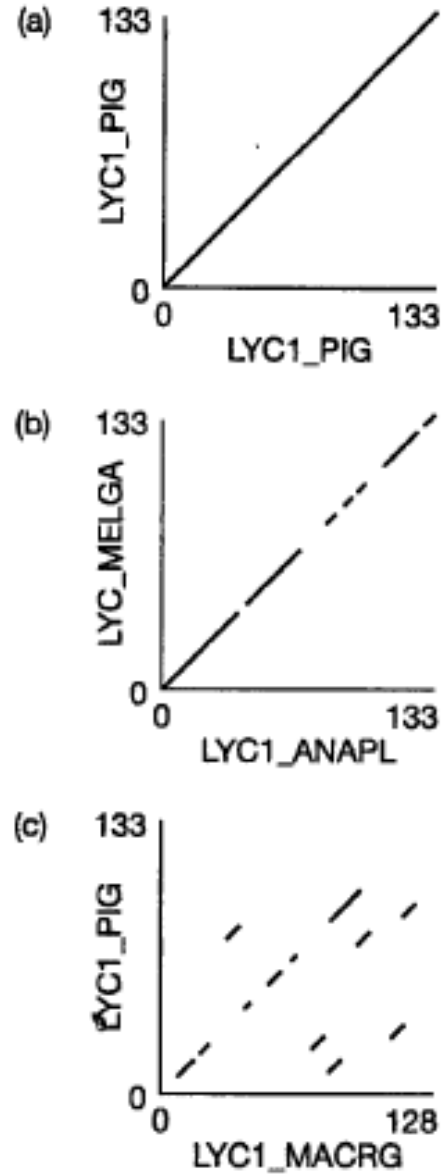
	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

(b)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	B	Z	X
C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Illustration of the manner of construction of the dotplot matrix, using a simple residue identity matrix to score an 'X' where a pair of identical residues is observed.

	M	T	F	R	D	L	L	S	V	S	F	E	G	P	R	P	D	S	S	A	G	G	S	S	A	G	G
M	X																										
T		X																									
F			X								X																
R				X											X												
D					X												X										
L						X	X																				
L						X	X																				
S								X		X								X	X				X	X			
V									X																		
S								X		X								X	X				X	X			
F			X								X																
E												X															
G													X								X	X				X	X
P														X		X											
R				X											X												
P														X		X											
D					X												X										
S								X		X								X	X				X	X			
S								X		X								X	X				X	X			
A																				X					X		
G													X								X	X				X	X
G													X								X	X				X	X



Graphical representation of dotplots, showing comparisons of (a) two identical sequences; (b) two highly similar sequences; and (c) two different, but related sequences.

Identity matrix used to initiate the Needleman and Wunsch alignment procedure for the sequences shown. Identical residue pairs are scored 1 in the appropriate matrix cell.

	A	D	L	G	A	V	F	A	L	C	D	R	Y	F	Q
A	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0
D	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
L	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
G	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
D	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Partially complete maximum-match matrix. The operation is about to be carried out on the highlighted cell L-L, by adding the highest value (5 at position C-C) from the two sub-paths leading to it, to give a value of 6.

	A	D	L	G	A	V	F	A	L	C	D	R	Y	F	Q
A	1	0	0	0	1	0	0	1	0	4	3	2	1	1	0
D	0	1	0	0	0	0	0	0	0	4	4	2	1	1	0
L	0	0	1	0	0	0	0	0	1	4	3	2	1	1	0
G	0	0	0	1	0	0	0	0	5	4	3	2	1	1	0
R	0	0	0	0	0	0	0	0	5	4	3	3	1	1	0
T	0	0	0	0	0	0	0	0	5	4	3	2	1	1	0
Q	0	0	0	0	0	0	0	0	5	4	3	2	1	1	1
N	0	0	0	0	0	0	0	0	5	4	3	2	1	1	0
C	0	0	0	0	0	0	0	0	4	5	3	2	1	1	0
D	0	1	0	0	0	0	0	0	3	3	4	2	1	1	0
R	0	0	0	0	0	0	0	0	2	2	2	3	1	1	0
Y	0	0	0	0	0	0	0	0	2	2	2	2	2	1	0
Y	0	0	0	0	0	0	0	0	1	1	1	1	2	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 6.6 Completed matrix in which the value being calculated in Table 6.5 is boxed, and the maximum-match pathway giving the highest scoring alignment is highlighted.

	A	D	L	G	A	V	F	A	L	C	D	R	Y	F	Q
A	9	7	6	6	7	6	6	7	5	4	3	2	1	1	0
D	7	8	6	6	6	6	6	6	5	4	4	2	1	1	0
L	6	6	7	5	5	5	5	5	**6**	4	3	2	1	1	0
G	5	5	5	6	5	5	5	5	5	4	3	2	1	1	0
R	5	5	5	5	5	5	5	5	5	4	3	3	1	1	0
T	5	5	5	5	5	5	5	5	5	4	3	2	1	1	0
Q	5	5	5	5	5	5	5	5	5	4	3	2	1	1	1
N	5	5	5	5	5	5	5	5	5	4	3	2	1	1	0
C	4	4	4	4	4	4	4	4	4	5	3	2	1	1	0
D	3	4	3	3	3	3	3	3	3	3	4	2	1	1	0
R	2	2	2	2	2	2	2	2	2	2	2	3	1	1	0
Y	2	2	2	2	2	2	2	2	2	2	2	2	2	1	0
Y	1	1	1	1	1	1	1	1	1	1	1	1	2	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

```
ADLGAVFALCDRYFQ
||||     ||||  |
ADLGRTQN-CDRYYQ
```

Figure 6.7 Final gapped alignment resulting from an implementation of the Needleman and Wunsch algorithm.

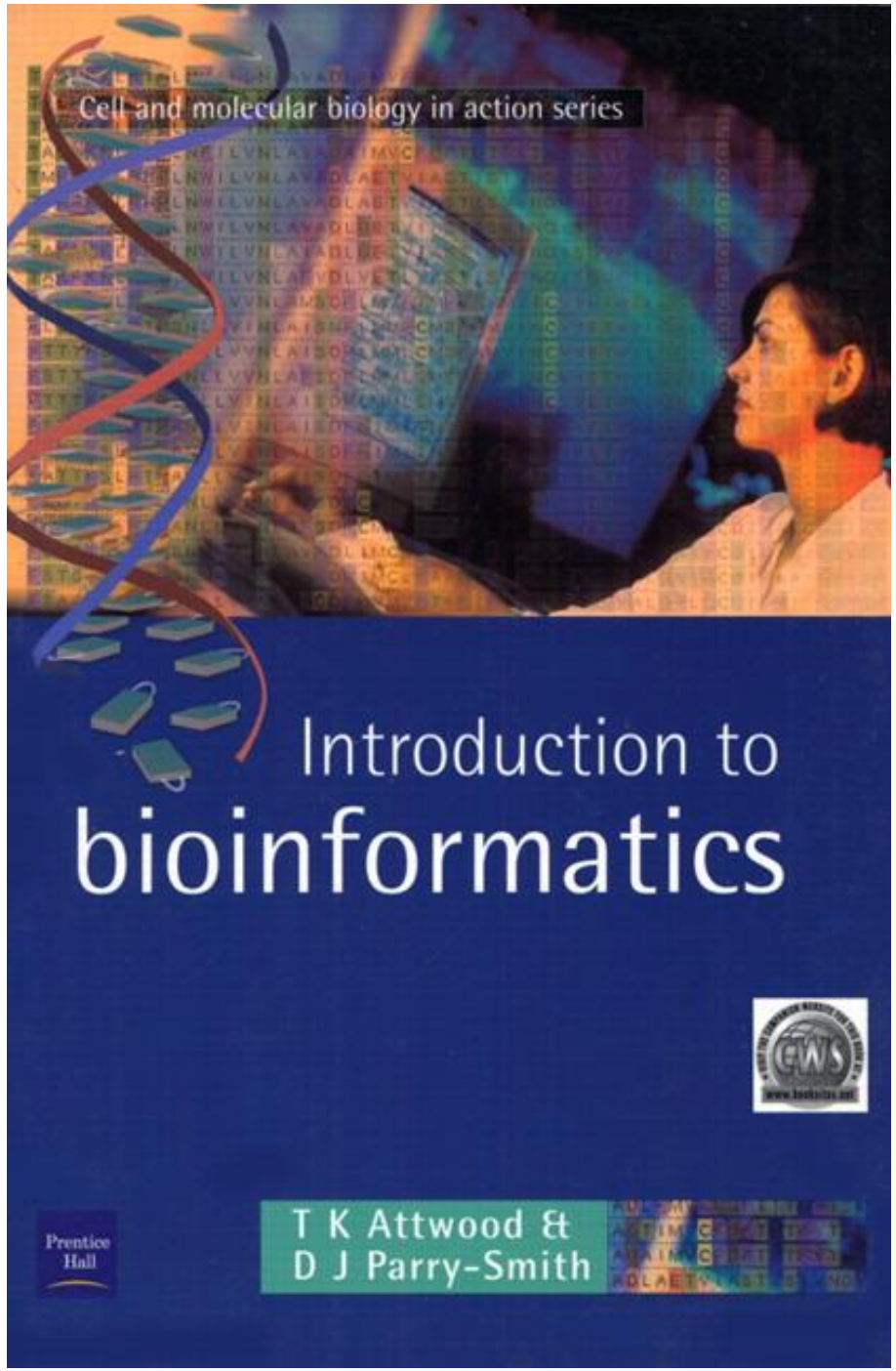
```
FASTA version 3.0t82 November 1, 1997
Please cite: W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

>gi|631066|pir||JC2331 adrenergic receptor alpha 1A - human, 572 bases
 vs SWISS-PROT Protein Sequence Database (rel35) library
25083768 residues in 69113 sequences
 statistics extrapolated from 50000 to 68413 sequences
 Expectation_n fit: rho(ln(x))= 6.3487+/-0.000531; mu= 6.8138+/- 0.030;
 mean_var=205.1722+/-43.131, 2-trim: 515 B-trim: 2588 in 1/63

FASTA (3.08 July, 1997) function (optimized, blosum matrix) ktup: 2
 join: 37, opt: 25, gap-pen: -12/ -2, width: 16 reg.-scaled
 Scan time: 12.420
The best scores are:                                      initn initl opt z-sc E(68413)
SW:A1AA_HUMAN P25100 homo sapiens (human ( 572) 3836 3836 3836 2695.2 1.8e-143
SW:A1AA_RAT P23944 rattus norvegicus (ra ( 561) 2691 2259 3156 2220.5 4.9e-117
SW:A1AB_RAT P15823 rattus norvegicus (ra ( 515) 1618 1019 1617 1146.5 3.2e-57
SW:A1AB_HUMAN P35368 homo sapiens (human ( 519) 1620 1011 1615 1145.0 3.9e-57
SW:A1AB_MESAU P18841 mesocricetus auratu ( 515) 1618 1019 1608 1140.2 7.3e-57
SW:A1AC_HUMAN P35348 homo sapiens (human ( 466) 1423  935 1464 1040.1 2.7e-51
SW:A1AC_RAT P43140 rattus norvegicus (ra ( 466) 1439  933 1458 1035.9 4.7e-51
SW:A1AC_BOVIN P18130 bos taurus (bovine) ( 466) 1417  922 1443 1025.4 1.8e-50
SW:A1AA_ORYLA Q91175 oryzias latipes (me ( 470) 1413  956 1434 1019.1 4e-50
SW:A1AB_CANFA P11615 canis familiaris (d ( 417) 1372  772 1366 972.2 1.7e-47

-----

>>SW:A1AA_RAT P23944 rattus norvegicus (rat). alpha-1a a (561 aa)
 initn: 2691 init1: 2259 opt: 3156 Z-score: 2220.5 expect() 4.9e-117
Smith-Waterman score: 3156; 85.315% identity in 572 aa overlap

               10        20        30        40        50        60
gi|631 MTFRDLLSVSFEGPRPDSSAGGSSAGGGGGGAGGAAPSEGPAVGGVPGGAGGGGGVVGAG
       :::::.:::.::::: .::.:::.::::.: .:   : :: ::::::: : :::.:::.:
SW:A1A MTFRDILSVTFEGPRSSSSTGGSGAGGGAGTVG---P-EGGAVGGVPG-ATGGGAVVGTG
               10        20        30            40          50

               70        80        90       100       110       120
gi|631 SGEDNRSSAGEPGSAGAGGDVNGTAAVGGLVVSAQGVGVGVFLAAFILMAVAGNLLVILS
       :::::.::.::::.: :.:.:::.:::::::::::::::::::::::::: ::::::::::
SW:A1A SGEDNQSSTGEPGAA-ASGEVNGSAAVGGLVVSAQGVGVGVFLAAFILTAVAGNLLVILS
           60        70        80        90       100       110

              130       140       150       160       170       180
gi|631 VACNRHLQTVTNYFIVNLAVADLLLSATVLPFSATMEVLGFWAFGRAFCDVWAAVDVLCC
       ::::::::::::::::::::::::::::.:::::::::::::::::::.:::::::::::::
SW:A1A VACNRHLQTVTNYFIVNLAVADLLLSAAVLPFSATMEVLGFWAPGRTFCDVWAAVDVLCC
          120       130       140       150       160       170

              190       200       210       220       230       240
gi|631 TASILSLCTISVDRYVGVRHSLKYPAIMTERKAAAILALLWVVALVVSVGPLLGWKEPVP
       ::::::::::::::::::::::::::::::::::::::::::.::::::::::::::::::
SW:A1A TASILSLCTISVDRYVGVRHSLKYPAIMTERKAAAILALLWAVALVVSVGPLLGWKEPVP
          180       190       200       210       220       230

-----

              490       500       510       520       530       540
gi|631 QAPVASRRKPPSAPREWRLLGPFRRPTTQLRAKVSSLSHKIPAGGAQRAEAACAQRSEVE
       :  :.: ::: ::.:::::::::..:::::::::::::::::::: .: :.:::.::: :::::
SW:A1A QDSVSSSRKPASALREWRLLGPLQRPTTQLRAKVSSLSHKIRSG-ARRAETACALRSEVE
              480       490       500       510       520

              550       560       570
gi|631 AVSLGVPHEVAEGATCQAYELADYSNLRETDI
       ::::.::.. ::.. ::::: .:::::::::::
SW:A1A AVSLNVPQDGAEAVICQAYEPGDYSNLRETDI
    530       540       550       560

-----
```

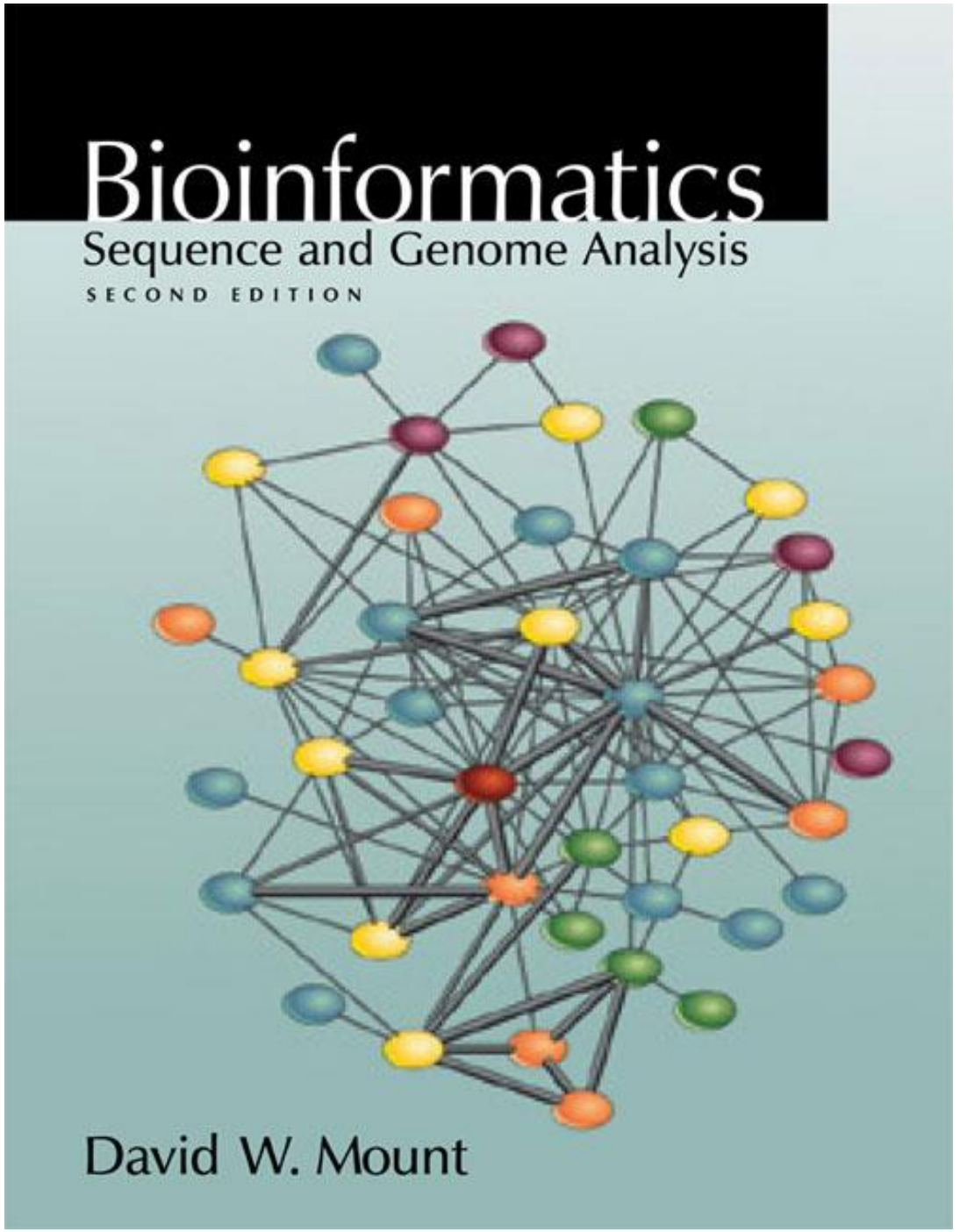
Excerpt from a typical FastA output (----- denotes excised material).

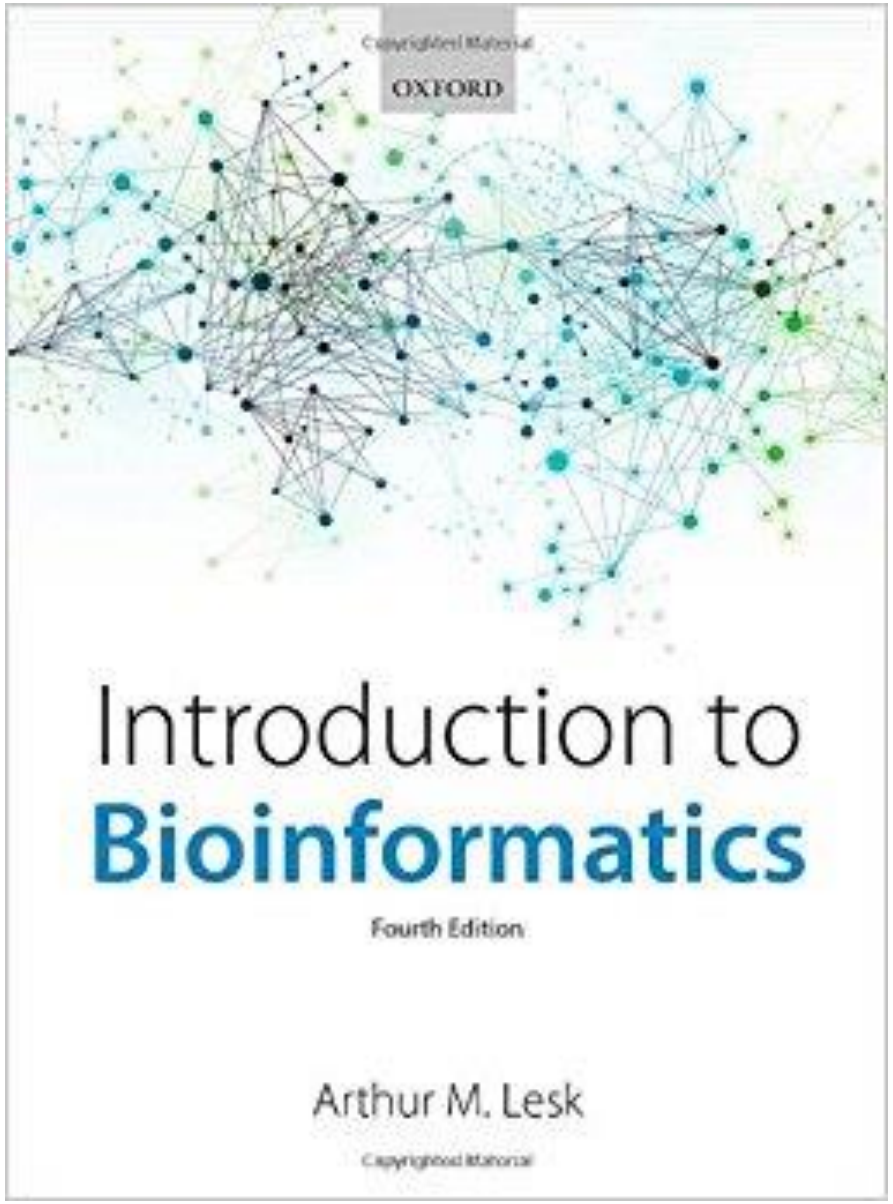
```
BLASTP 1.4.11 [24-Nov-97] [Build 24-Nov-97]

Reference: Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David
J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-10.
Query= gi|631066|pir||JC2331 adrenergic receptor alpha 1A - human (572 letters)
Database: Non-redundant SwissProt (74,037 sequences; 26,661,674 total letters)
Searching.....................................................done

                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)       N

sp|P25100|A1AD_HUMAN ALPHA-1D ADRENERGIC RECEPTOR (ALPHA ... 1513  5.5e-266  4
sp|O02666|A1AD_RABIT ALPHA-1D ADRENERGIC RECEPTOR (ALPHA ... 1465  3.9e-242  4
sp|P23944|A1AD_RAT   ALPHA-1D ADRENERGIC RECEPTOR (ALPHA ... 1416  2.0e-228  5
sp|P97714|A1AD_MOUSE ALPHA-1D ADRENERGIC RECEPTOR (ALPHA ... 1411  5.1e-220  3
sp|P15823|A1AB_RAT   ALPHA-1B ADRENERGIC RECEPTOR (ALPHA ...  650  9.2e-130  2
sp|P18841|A1AB_MESAU ALPHA-1B ADRENERGIC RECEPTOR (ALPHA ...  650  9.2e-130  2
sp|P35368|A1AB_HUMAN ALPHA-1B ADRENERGIC RECEPTOR (ALPHA ...  643  8.8e-129  2
sp|P97717|A1AB_MOUSE ALPHA-1B ADRENERGIC RECEPTOR (ALPHA ...  629  8.2e-127  2
sp|P35348|A1AA_HUMAN ALPHA-1A ADRENERGIC RECEPTOR (ALPHA ...  589  4.2e-118  2
sp|O02824|A1AA_RABIT ALPHA-1A ADRENERGIC RECEPTOR (ALPHA ...  591  1.1e-117  2

-----

sp|P25100|A1AD_HUMAN ALPHA-1D ADRENERGIC RECEPTOR (ALPHA 1D-ADRENOCEPTOR) Length = 572

 Score = 89 (41.7 bits), Expect = 5.5e-266, Sum P(4) = 5.5e-266
 Identities = 17/17 (100%), Positives = 17/17 (100%)

Query:     1 MTFRDLLSVSFEGPRPD 17
             MTFRDLLSVSFEGPRPD
Sbjct:     1 MTFRDLLSVSFEGPRPD 17

 Score = 1513 (708.4 bits), Expect = 5.5e-266, Sum P(4) = 5.5e-266
 Identities = 299/348 (85%), Positives = 299/348 (85%)

Query:    63 EDNRXXXXXXXXXXXXXDVNGTAAVGGLVVSAQGVGVGVFLAAFILMAVAGNLLVILSVA 122
             EDNR             DVNGTAAVGGLVVSAQGVGVGVFLAAFILMAVAGNLLVILSVA
Sbjct:    63 EDNRSSAGEPGSAGAGGDVNGTAAVGGLVVSAQGVGVGVFLAAFILMAVAGNLLVILSVA 122

Query:   123 CNRHLQTVTNYFIVNLAVADLLLSATVLPFSATMEVLGFWAPGRAFCDVWAAVDVLCCTA 182
             CNRHLQTVTNYFIVNLAVADLLLSATVLPFSATMEVLGFWAPGRAFCDVWAAVDVLCCTA
Sbjct:   123 CNRHLQTVTNYFIVNLAVADLLLSATVLPFSATMEVLGFWAPGRAFCDVWAAVDVLCCTA 182

Query:   183 SILSLCTISVDRYVGVRHSLKYPAIMTERKXXXXXXXXXXXXXXXXXXXXGWKEPVPPD 242
             SILSLCTISVDRYVGVRHSLKYPAIMTERK                    GWKEPVPPD
Sbjct:   183 SILSLCTISVDRYVGVRHSLKYPAIMTERKAAAILALLWVVALVVSVGPLLGWKEPVPPD 242

Query:   243 ERPCGITEEAGYAVFSSVCSFYLPMXXXXXXXXXXXXXXSTTRSLEAGVKRERGKASEV 302
             ERPCGITEEAGYAVFSSVCSFYLPM              STTRSLEAGVKRERGKASEV
Sbjct:   243 ERPCGITEEAGYAVFSSVCSFYLPMAVIVVMYCRVYVVARSTTRSLEAGVKRERGKASEV 302

Query:   303 VLRIHCRGAATGADGAHGMRSAKGHTFRSSLSVRLLKFSREKKAAKTLAIVVGVFVLCWF 362
             VLRIHCRGAATGADGAHGMRSAKGHTFRSSLSVRLLKFSREKKAAKTLAIVVGVFVLCWF
Sbjct:   303 VLRIHCRGAATGADGAHGMRSAKGHTFRSSLSVRLLKFSREKKAAKTLAIVVGVFVLCWF 362

Query:   363 PFFFVLPLGSLFPQLKPSEGVPKVIFWLGYFNSCVNPLIYPCSSREFK 410
             PFFFVLPLGSLFPQLKPSEGVPKVIFWLGYFNSCVNPLIYPCSSREFK
Sbjct:   363 PFFFVLPLGSLFPQLKPSEGVPKVIFWLGYFNSCVNPLIYPCSSREFK 410

 Score = 101 (47.3 bits), Expect = 5.5e-266, Sum P(4) = 5.5e-266
 Identities = 17/17 (100%), Positives = 17/17 (100%)

Query:   433 VYGHHWRASTSGLRQDC 449
             VYGHHWRASTSGLRQDC
Sbjct:   433 VYGHHWRASTSGLRQDC 449

 Score = 387 (181.2 bits), Expect = 5.5e-266, Sum P(4) = 5.5e-266
 Identities = 78/93 (83%), Positives = 78/93 (83%)

Query:   480 MQAPVASRRKPPSAFREWRLLGPFRRPTTQLRAKVSSLSHKIPXXXXXXXXXXXXXSEV 539
             MQAPVASRRKPPSAFREWRLLGPFRRPTTQLRAKVSSLSHKI                SEV
Sbjct:   480 MQAPVASRRKPPSAFREWRLLGPFRRPTTQLRAKVSSLSHKIRAGGAQRAEAACAQRSEV 539

Query:   540 EAVSLGVPHEVAEGATCQAYELADYSNLRETDI 572
             EAVSLGVPHEVAEGATCQAYELADYSNLRETDI
Sbjct:   540 EAVSLGVPHEVAEGATCQAYELADYSNLRETDI 572

-----
```

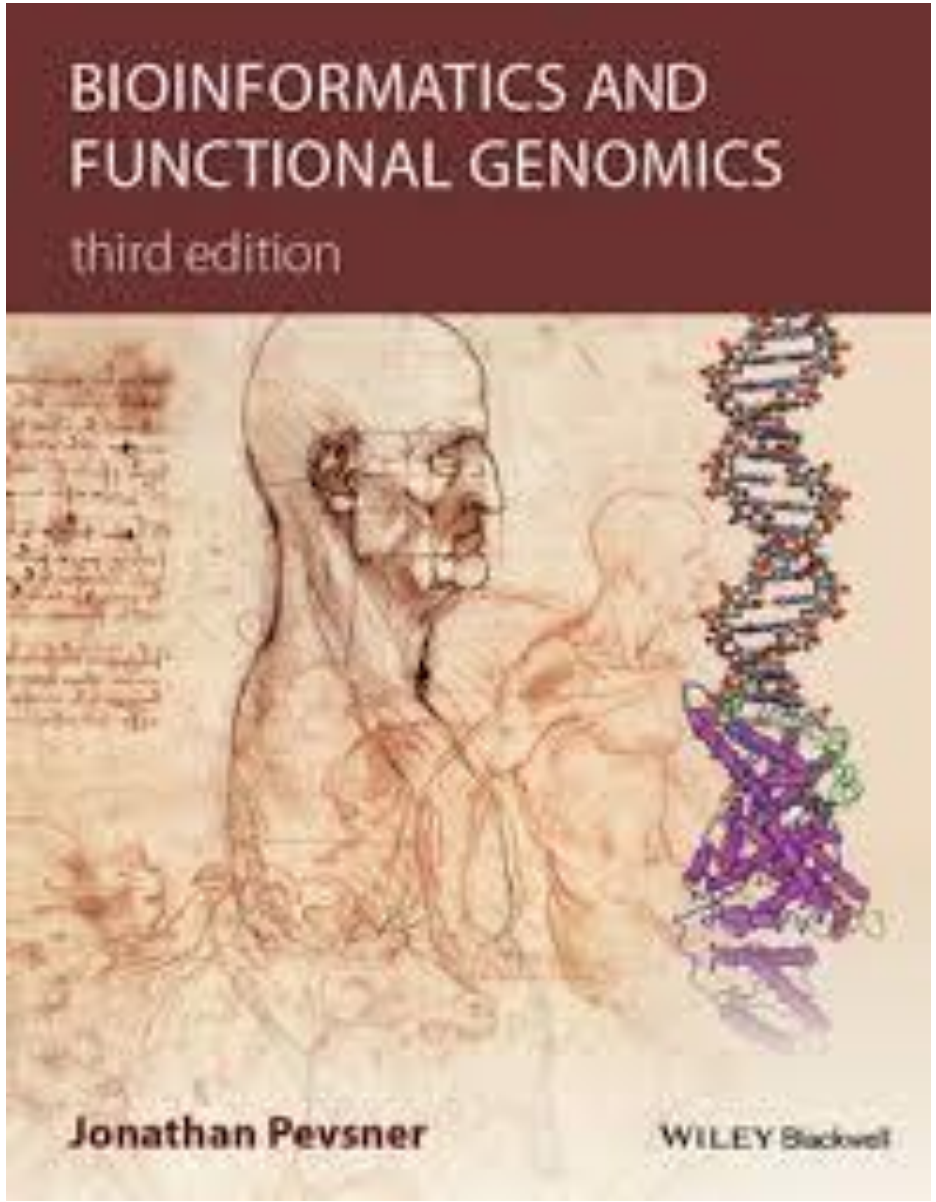
Excerpt from a typical BLAST output (----- denotes excised material).

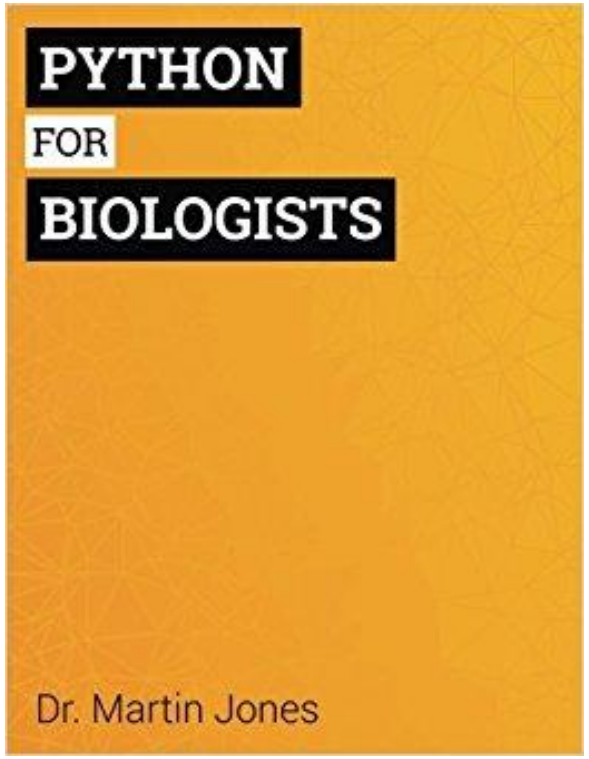
REFERENCES

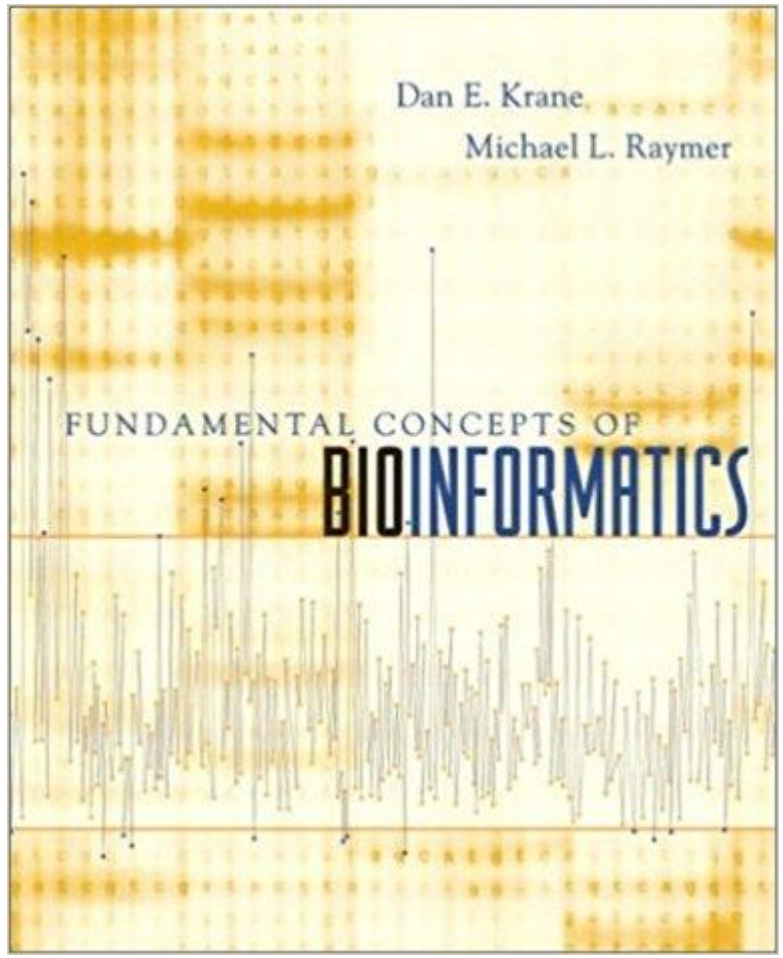

Cell and molecular biology in action series
Introduction to
bioinformatics
T K Attwood &
D J Parry-Smith
Prentice
Hall

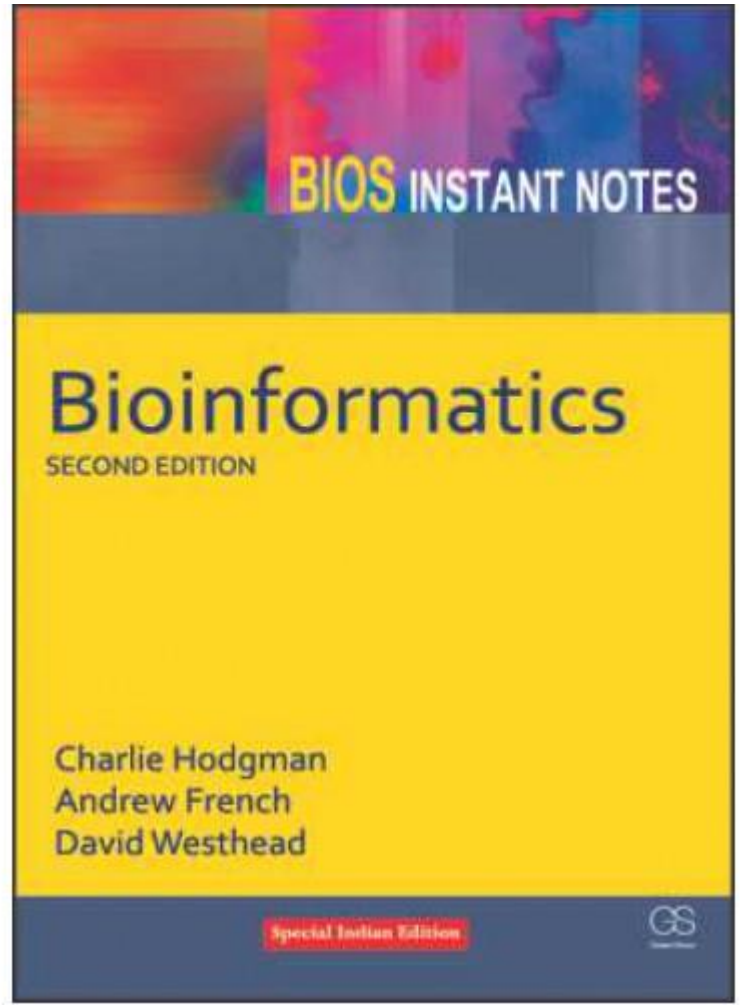

Bioinformatics
Sequence and Genome Analysis
SECOND EDITION
David W. Mount


OXFORD
Introduction to
Bioinformatics
Fourth Edition
Arthur M. Lesk


BIOINFORMATICS AND
FUNCTIONAL GENOMICS
third edition
Jonathan Pevsner
WILEY Blackwell


PYTHON
FOR
BIOLOGISTS
Dr. Martin Jones


Dan E. Krane
Michael L. Raymer
FUNDAMENTAL CONCEPTS OF
BIOINFORMATICS


BIOS INSTANT NOTES
Bioinformatics
SECOND EDITION
Charlie Hodgman
Andrew French
David Westhead
Special Indian Edition
GS

Bioinformatics is the future…