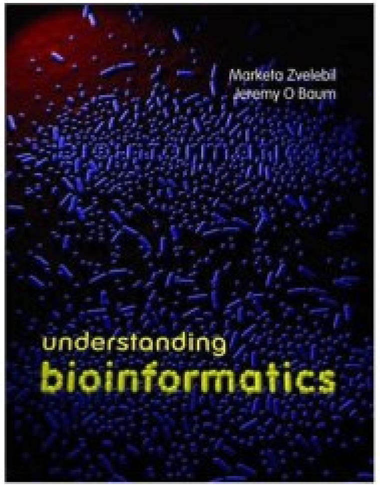
# Useful Web Sites for Part 2
# Sequence Alignments



## Chapter 4: Producing and Analyzing Sequence Alignments
## Chapter 5: Pairwise Sequence Alignment and Database Searching
## Chapter 6: Patterns, Profiles and Multiple Alignments

**Sequence formats:**

There are many different ways of representing individual sequences and multiple aligned sequences in text files.

An excellent introductory guide with examples is available on an EBI help page

The EMBOSS suite of analysis programs also contains a guide to alternative formats and their inter-conversion

Many sequence analysis programs can read and write several formats, but occasionally it is necessary to convert between formats in order to use an application. Two free programs available are:

Readseq – web server with download link
SeqVerter

**Sequence and sequence alignment databases:**

The following are the sites with general information about specific databases:

Nucleotide sequence databases:

GenBank is a sequence database, including specialised sections such as dbEST for expressed sequence tags.

EMBL Nucleotide Sequence database

Protein sequence databases:

UniProt is a protein sequence resource, which contains specialised sections including UniProtKB/Swiss-Prot (also known as Swiss-Prot) and UniProtKB/TrEMBL (also known as TrEMBL)

Sequence pattern and motif databases:

- BLOCKS database of aligned protein sequences

Domain and protein family databases:

- InterPro database of aligned sequences of protein families, domains and functional sites
- Prodom comprehensive set of protein domain families as aligned sequences

Protein sequence patterns databases:

- Prosite

Protein sequence profile HMM databases:

- Pfam

Multiple alignment databases:

- FSSP (families of structurally similar proteins)
- Homologous Structure Alignment Database (HOMSTRAD)

To access the data, use either of the following sites:

- European Bioinformatics Institute (EBI)
- National Center for Biotechnology Information (NCBI) Tools for Data Mining

## Other useful information:

Coping with limited data:

- Dirichlet mixtures and other prior distributions

Multiple alignment test databases:

- BAliBASE versions 1 and 2; version 3
- Oxbench
- Protein Reference Alignment Benchmark (PREFAB)

## Programs

Sequence format conversion programs:

- SeqVerter™

Database search and pairwise alignment programs:

- Dotter
- FASTA
- NCBI BLAST which includes PSI-BLAST

FSA-BLAST

WU BLAST 2.0

Low complexity sequence mask programs:

DUST and DustMasker

SEG

Very long sequence and genome alignment programs:

BLASTZ

BLAT

CHAOS

MUMMER3

LAGAN and associated programs

SSAHA

Multiple sequence alignment programs:

ClustalW

DIALIGN

MAFFT

MSA

MUSCLE

ProbCons

SAGA

SATCHMO is implemented in LOBSTER

T-COFFEE

An extensive listing is available at http://en.wikipedia.org/wiki/Sequence_alignment_software.

Hidden Markov model programs:

Sequence Alignment and Modeling System (SAM)

HMMER

Alignment visualisation and formatting programs:

CINEMA

PFAAT

WebLogo

Programs for aligning multiple alignments:

prof_sim

COMPASS

Programs for aligning HMMS:

COACH is implemented in LOBSTER
HHsearch

Programs for identifying common patterns in a set of sequences:

- Gibbs
- MEME
- PRATT

Programs for finding known patterns in a sequence:

- MAST
- ps_scan is a perl script to search for PROSITE patterns

## Web servers

Apart from the web pages at the major bioinformatics resource sites listed on the web page for Part 1, which offer access to many alignment programs, the following web sites provide on-line access to sequence analysis programs:

- William R. Pearson's FASTA programs at the University of Virginia
- PRRN
- WebLogo

## Datafiles used for Chapter 4 examples

Files will be found in the archive 'Part 2 Sequence Alignments datafiles.zip'.

| | | |
|---|---|---|
| Breast cancer susceptibility gene protein BRCA2 sequences | Fig 4.3 | BRCA2.seq |
| A cAMP-dependent protein kinase and related PI3-kinase p110 sequences | Fig 4.5, 4.7, 4.10, 4.12-4.13, 4.15 | cAMPKinase.seq,<br>MultipleKinaseSequences.seq |
| Five SH2 domains | Fig 4.11 | SH2domain.seq |
| Human prion precursor protein (PrP) | Fig 4.14, 4.18 | PrionProtein.seq |