

B.Sc. (Honours) Botany & Microbiology  
**Semister V**  
Discipline Specific Elective- DSE II  
Choice Based Credit System (CBCS)

# Bioinformatics

Dr. Pritam Chattopadhyay  
Assistant Professor  
Department of Botany  
M.U.C. Women's College, Burdwan

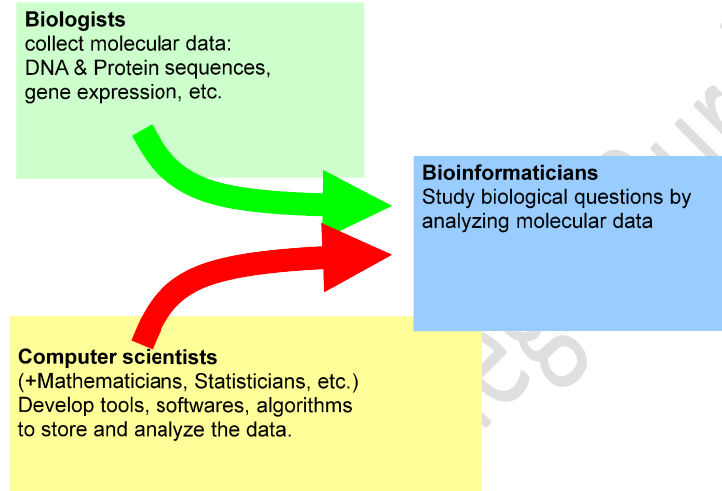
## Unit 1. Introduction to Bioinformatics

### Introduction

#### What is bioinformatics?

Oxford English Dictionary defines bioinformatics as: “The branch of science concerned with information and information flow in biological systems, esp. the use of computational methods in genetics and genomics”

The National Center for Biotechnology Information (NCBI) defines bioinformatics as: "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline"



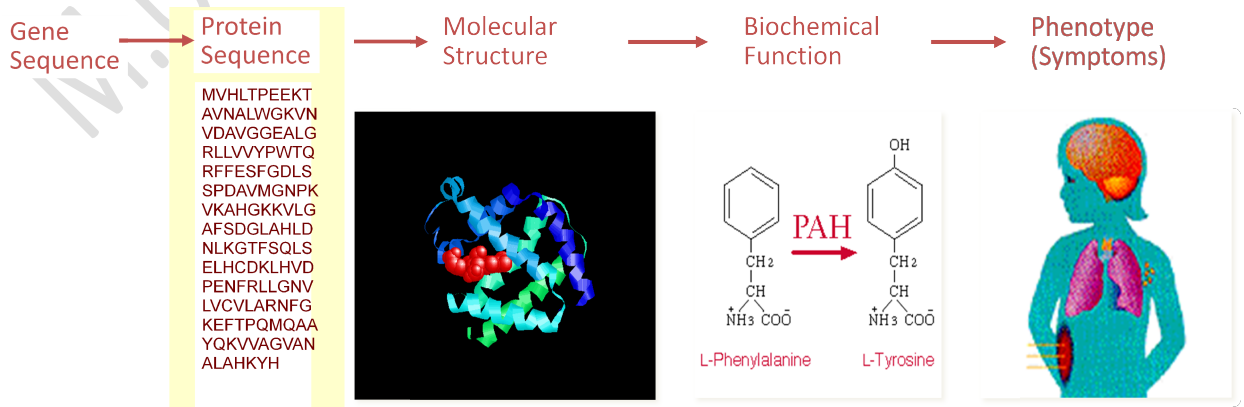
#### Central Paradigm of Molecular Biology

- Flow of biological information within cell

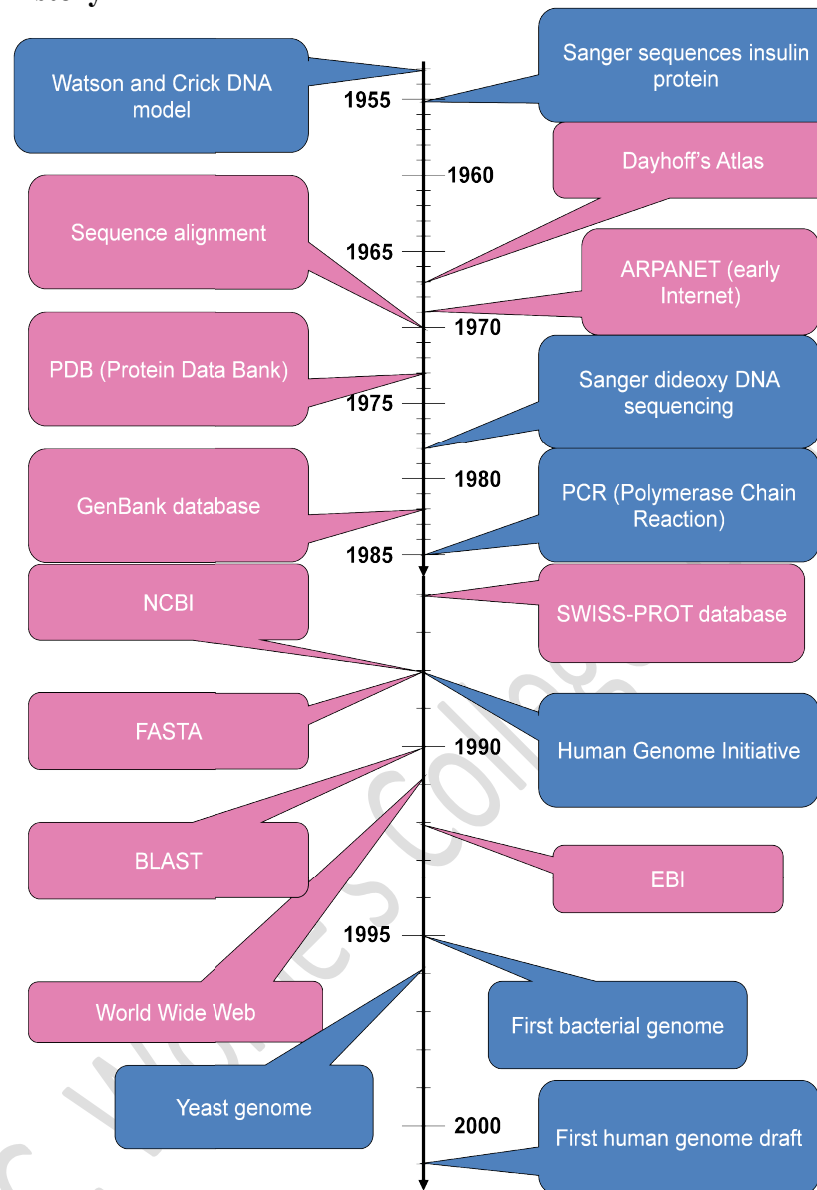


#### Central Paradigm of Bioinformatics

- Genetic information is redundant
- Structural information is redundant
- Genes and proteins are meta-stable
- Single genes have multiple functions
- Genes are one dimensional but function depends on three-dimensional structure

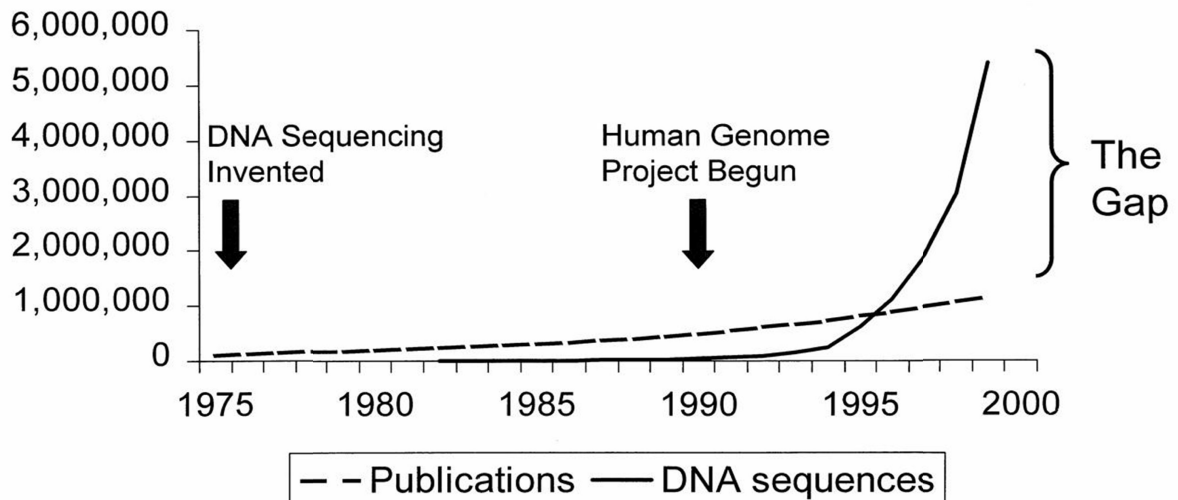


## Bioinformatics history



- The first protein sequence reported was that of bovine insulin in 1956, consisting of 51 residues by **Fred Sanger**
- Nearly a decade later, the first nucleic acid sequence was reported, that of yeast tRNA<sup>alanine</sup> with 77 bases
- First complete viral genome 1978 by **Fred Sanger**
- The birth of bioinformatics: In 1965, **Margaret Oakley Dayhoff** gathered all the available sequence data to create the first bioinformatic database (*Atlas of Protein Sequence and Structure*)
  - The first protein database
  - The first program for sequence assembly
- The Protein DataBank followed in 1972 with a collection of ten X-ray crystallographic protein structures.
- The SWISSPROT protein sequence database began in 1987

- 1989 saw the introduction of the microarray technique by Stephen Fodor
- Term “Bioinformatics” first came into use in the 1990s
- The first high-throughput genomics technology was automated DNA sequencing in the early 1990
- In 1995, Venter and Hamilton used whole-genome shotgun sequencing strategy to sequence the genomes of *Mycoplasma* and *Haemophilus*



(Boguski, 1999)

- From 1995, the number of genes in the database started to exceed the number of papers on molecular biology and genetics in the literature
- Because it is impossible to maintain up-to-date copies of all relevant databases within the lab. Access to the data is via the internet
- In September 1999, Celera Genomics completed the sequencing of the *Drosophila* genome
- In 2001, the 3-billion-bp human genome sequence was generated in a competition between the publicly funded Human Genome Project and Celera

### Components of Bioinformatics

(Key skills required for Bioinformatics)

- Strong background in some aspect of molecular biology
- Ability to communicate biological questions comprehensibly to computer scientists
- Thorough comprehension of the problem in the bioinformatics field
- Statistics (association studies, clustering, sampling)
- Ability to filter, parse, and munge data and determine the relationships between the data sets
- Mathematics (e.g. algorithm development)
- Engineering (e.g. robotics)
- Good knowledge of a few molecular biology software packages (molecular modeling / sequence analysis)
- Command line computing environment (Linux/Unix knowledge)
- Data administration (esp. relational database concept) and Computer Programming Skills/Experience (C/C++, Sybase, Java, Oracle) and Scripting Language Knowledge (Perl and Python)

## **Branches of Bioinformatics**

### **1. Structural Bioinformatics**

Prediction of macromolecular 3D structures from their respective sequences, understanding the protein folding and mis-folding mechanisms, which may lead to diseases.

### **2. Computational Biology**

Interdisciplinary field involving various concepts of applied mathematics, statistics, and programming languages to address the problems in Biology.

### **3. Drug Designing and Delivery**

Good understanding of the disease mechanism and pathogenicity & designing of drugs based on the biological targets.

### **4. Analysis of Next Generation Sequencing (NGS)**

Handling NGS data including miRNA has great implications in many aspects including diseases like cancer and is one of the emerging fields in Bioinformatics.

### **5. Phylogenetics**

Predicting the evolutionary relationships among a set of organisms is another important branch of bioinformatics.

### **6. Software and Database Development**

There are various bioinformatic softwares used to predict the results of wet lab experiments like structure determination by simulation of macromolecules which effectively save time. Also efficient maintenance of macromolecular databases is another important part of bioinformatics.

## **Aim, Scope and Research Areas of Bioinformatics**

### **Aim of Bioinformatics**

(What do we actually do with bioinformatics?)

- Sequence assembly
- Genome annotation
- Molecular evolution
- Analysis of gene expression
- Analysis of gene regulation
- Protein structure prediction
- Protein docking

### **Scopes of Bioinformatics**

(Applications of Bioinformatics)

- Novel Diagnostics
  - Microchips & Microarrays - DNA
  - Gene Expression - RNA
  - Proteomics - Protein
- Novel Therapeutics
  - Drug Target Discovery
  - Rational Drug Design
  - Molecular Docking
  - Gene Therapy
  - Stem Cell Therapy

- Understanding Metabolism
- Crop Improvement
- Understanding Disease
  - Inherited Diseases - OMIM
  - Infectious Diseases
  - Pathogenic Bacteria
  - Viruses
- Bioinformatics market
  - Current is worth 300 million / year
  - Prediction: \$2 billion / year in 5-6 years
  - Major players (Bioinformatics companies) - Genomatrix Software, Genaissance Pharmaceuticals, Lynx, Lexicon Genetics, DeCode Genetics, CuraGen, AlphaGene, Bionavigation, Pangene, InforMax, TimeLogic, GeneCodes, LabOnWeb.com, Darwin, Celera, Incyte, BioResearch Online, BioTools, Oxford Molecular, Genomica, NetGenics, Rosetta, Lion BioScience, DoubleTwist, eBioinformatics, Prospect Genomics, Neomorphic, Molecular Mining, *GeneLogic*, *GeneFormatics*, *Molecular Simulations*, *Bioinformatics Solutions* etc.

### **Research Areas of Bioinformatics**

“Omics” projects and the information associated with involve a huge amount of data that is stored on computers all over the world. 21<sup>st</sup> century is a century of biotechnology and OMICS.

- Genomics: New sequence information is being produced at increasing rates. (The contents of GenBank double every year)
- Transcriptomics: RNA levels of every gene in the genome analyzed in parallel. Microarray: Global expression analysis Progressively replaced by RNA-seq
- Proteomics: Global protein analysis generates by large massspectra libraries
- Metabolomics: Global metabolite analysis: 25,000 secondary metabolites characterized

## Unit 2. Databases in Bioinformatics

### Introduction

#### ➤ Database or databank?

Initially

- Databank (in UK)
- Database (in the USA)

Solution

- The abbreviation *db*

#### ➤ What is a Database?

A structured collection of data held in computer storage; *esp.* one that incorporates software to make it accessible in a variety of ways; *transf.*, any large collection of information.

- **Database management:** the organization and manipulation of data in a database.
- **Database management system (DBMS):** a software package that provides all the functions required for database management.
- **Database system:** a database together with a database management system.

#### ➤ Components of Database

- structured
- searchable (index) → table of contents
- updated periodically (release) → new edition
- cross-referenced (hyperlinks) → links with other db

#### ➤ Overview of Database

1. Database indexing and specification of search terms (retrieval, follow-up, analysis)
2. Archives (databases on: nucleic acid sequences, genome, protein sequences, structures, proteomics, expression, pathways)
3. Gateways to Archives (NCBI, Entrez, PubMed, ExPasy, Swiss-Prot, SRS, PIR, Ensembl)

### FASTA: a standard data format

Easy to manage: all the entries are visible at the same time !

Flat-file database (« flat file, 3 entries »):

```
Accession number: 1
First Name: Amos
Last Name: Bairoch
Course: Pottery 2000; Pottery 2001;
//
Accession number: 2
First Name: Dan
Last name: Graur
Course: Pottery 2000, Pottery 2001; Ballet 2001, Ballet 2002
//
Accession number 3:
First Name: John
Last name: Travolta
Course: Ballet 2001; Ballet 2002;
//
```

## Biological Databases

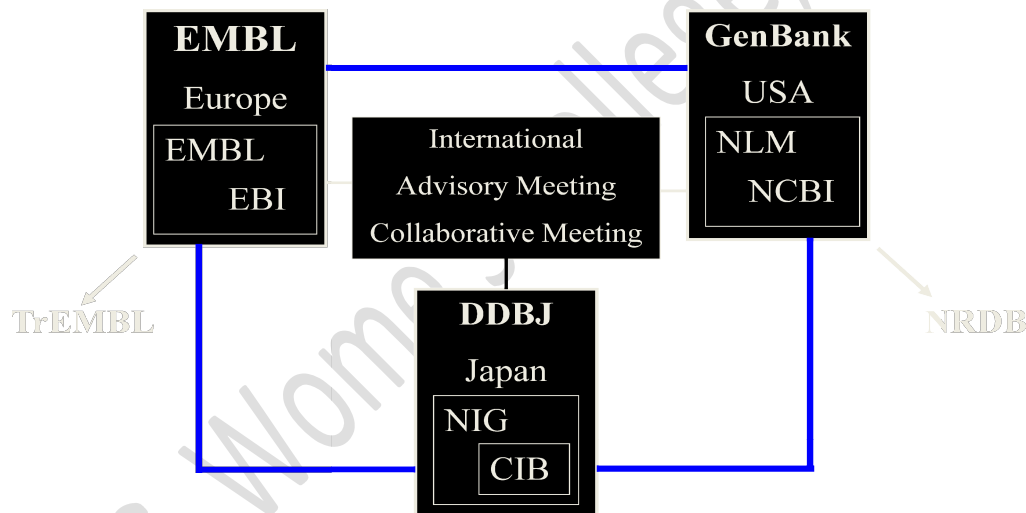
### ➤ Why biological databases?

- Exponential growth in biological data.
- Data (genomic sequences, 3D structures, 2D gel analysis, MS analysis, Microarrays....) are no longer published in a conventional manner, but directly submitted to databases.
- Essential tools for biological research. The only way to publish massive amounts of data without using all the paper in the world.

### ➤ Some statistics

- More than 1000 different 'biological' databases
- Variable size: <100Kb to >20Gb
  - DNA: > 20 Gb
  - Protein: 1 Gb
  - 3D structure: 5 Gb
  - Other: smaller
- Update frequency: **daily** to **annually** to **seldom** to **forget about it**.
- Usually accessible through the web (some free, some not)

### ➤ International nucleotide data banks



### ➤ Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolic networks
- Regulatory networks
- Bibliography
- Expression (Microarrays,...)
- Specialized

## Classification of Biological Databases

- Based on type of sequence Databases are of three types:
  - a) DNA (Nucleotide) sequences databases
    - GenBank (NCBI): <http://www.ncbi.nlm.nih.gov>
    - Ensembl: <http://useast.ensembl.org/index.html>
    - DDBJ (DNA DataBase of Japan): <http://www.ddbj.nig.ac.jp>
    - TIGR: <http://tigr.org/tdb/tgi>
  - b) Protein (amino acid) databases
    - Swiss-Prot (very high level of annotation) <http://au.expasy.org/>
    - PIR (protein identification resource) the world's most comprehensive catalog of information on proteins <http://www.pir.uniprot.org/>
    - PDB (sequences derived from the 3D structure Brookhaven PDB) <http://www.rcsb.org/pdb/>
  - c) Translated databases
    - TREMBL (translated EMBL): includes entries that have not been annotated yet into Swiss-Prot <http://www.ebi.ac.uk/trembl/access.html>
    - GenPept (translation of coding regions in GenBank)
- Based on Data types Databases are of four types:
  - a) Primary databases
    - Database containing primary data (sequence): EMBL (European Molecular Biology Laboratory nucleotide sequence database at EBI, Hinxton, UK); GenBank (at National Center for Biotechnology information, NCBI, Bethesda, MD, USA); DDBJ (DNA Data Bank Japan at CIB, Mishima, Japan)
  - b) Secondary database
    - Database containing secondary data (motifs): SP; PseqIP
  - c) Tertiary database
    - Database containing tertiary data (atomic co-ordinates): NRL3D; PIRMOD
  - d) Composite datyabase
    - Database containing more than one database: NRDB (PIR+SP+PDB+GenPept); OWL (PIR+SP+GeneBank+NRL3D); MIPSX (PIR+PATCHX)
- Based on purpose Databases are of two types:
  - a) Generalized Database
    - Generalized Protein sequence databases: SWISS-PROT; TrEMBL; PIR-PSD; UniProt; UniRef; IPI
    - Generalized Carbohydrate databases: CarbBank
    - Generalized 3D structure databases: PDB; EBI-MSD; NDB
  - b) Specialized Database
    - Databases for EST, STS, SNP, RNA, genomes, protein families, pathways, microarray data etc.

## Biological Database Retrieval System

Sequence Retrieval System (SRS)- Users can formulate queries across a range of different database types.

### ➤ Database search may be

- **Text-based** (SRS, Entrez ...)
- **Sequence-based** (sequence similarity search) (BLAST, FASTA...)
- **Motif-based** (ScanProsite, eMOTIF)
- **Structure-based** (structure similarity search) (VAST, DALI...)
- **Mass-based** protein search (ProteinProspector, PeptIdent, Prowl ...)

### ➤ The Syntax of a Query term

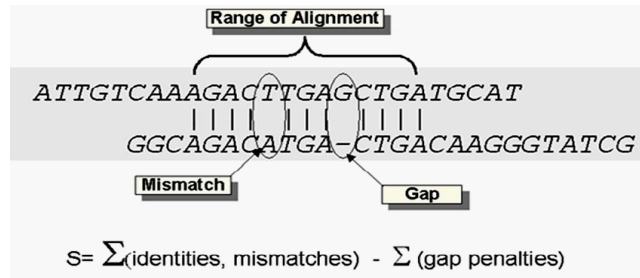
1. Boolean operators: AND, OR, NOT must be entered in UPPERCASE (e.g., promoters OR response elements). The default is AND.
2. Entrez processes all Boolean operators in a left-to-right sequence. The order in which Entrez processes a search statement can be changed by enclosing individual concepts in parentheses. The terms inside the parentheses are processed first. For example, the search statement: g1p3 OR (response AND element AND promoter).
3. Quotation marks: The term inside the quotation marks is read as one phrase (e.g. “public health” is different than public health, which will also include articles on public latrines and their effect on health workers).
4. Asterisk: Extends the search to all terms that start with the letters before the asterisk. For example, dia\* will include such terms as diaphragm, dial, and diameter.

### ➤ Query Refinement

- Often a search finds too many (or too few) sequences, so you can go back and try again with more (or fewer) keywords in your query
- The “History” feature allows you to combine any of your past queries.
- The “Limits” feature allows you to limit a query to specific organisms, sequences submitted during a specific period of time, etc.
- [Many other features are designed to search for literature in MEDLINE]

### ➤ Database homology searching

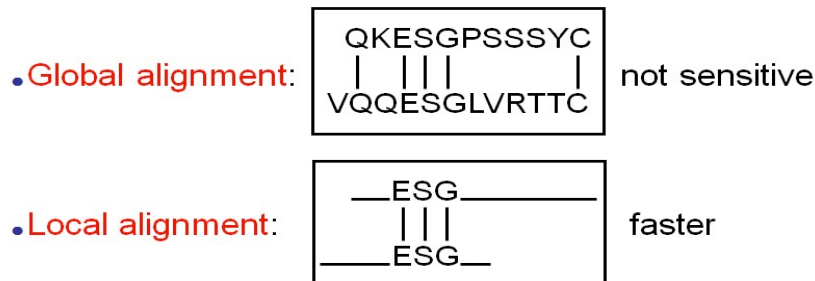
- Use algorithms to efficiently provide mathematical basis of searches that can be translated to statistical significance
- Assumes that sequence, structure, and function are inter-related
- All similarity searching methods rely on the concepts of alignment and distance between sequences
- A similarity score is calculated from a distance: the number of DNA bases or amino acids that are different between two sequences
- Scoring system: Uses scoring matrices that allow biologists to quantify the quality of sequence alignments



- The raw score  $S$  is calculated by summing the scores for each aligned position and the scores for gaps. Gap creation/extension scores are inherent to the scoring system in use (BLAST, FASTA...)
- The score for an identity or a mismatch is given by the specified substitution matrix (e.g., BLOSUM62)
- Some popular scoring matrices are:  
PAM (Percent Accepted Mutation): for evolutionary studies. For example in PAM1, 1 accepted point mutation per 100 amino acids is required  
BLOSUM (BLOCKs amino acid SUBstitution Matrix): for finding common motifs. For example in BLOSUM62, the alignment is created using sequences sharing no more than 62% identity
- Understanding theories underlying a given scoring matrix can aid in making proper choice:
  - Some matrices reflect similarity: good for database searching
  - Some reflect distance: good for phylogenies

➤ **Database search methods: Sequence Alignment**

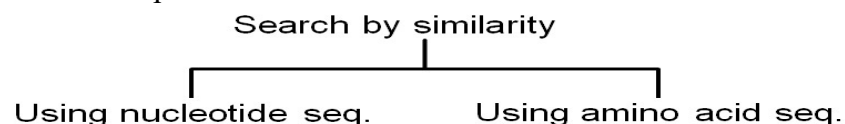
- Two broad classes of sequence alignments exist:



- The most widely used local similarity algorithms are:
  - Smith-Waterman (<http://www.ebi.ac.uk/MPsrch/>)
  - Basic Local Alignment Search Tool (BLAST, <http://www.ncbi.nih.gov>)
  - Fast Alignment (FASTA, <http://fasta.genome.jp>; <http://www.ebi.ac.uk/fasta33/>)

➤ **Tools to search databases**

- BLAST and FASTA
- The dilemma: DNA or protein?



- It is almost always better to compare coding sequences in their amino acid form, especially if they are very divergent

- Very highly similar nucleotide sequences may give better results
- Therefore there are variants of BLAST and FASTA
- **FASTA:** *Compares a DNA query to DNA database, or a protein query to protein database*
- **FASTX:** *Compares a translated DNA query to a protein database*
- **TFASTA:** *Compares a protein query to a translated DNA database*
- **BLASTN:** *Compares a DNA query to DNA database.*
- **BLASTP:** *Compares a protein query to protein database.*
- **BLASTX:** *Compares the 6-frame translations of DNA query to protein database.*
- **TBLASTN:** *Compares a protein query to the 6-frame translations of a DNA database. You can however define your frame of interest*
- **TBLASTX:** *Compares the 6-frame translations of DNA query to the 6-frame translations of a DNA database (each sequence is comparable to BLASTP searches!)*
- **PSI-BLAST:** *Performs iterative database searches. The results from each round are incorporated into a 'position specific' score matrix, which is used for further searching*

#### ➤ Database searching strategies

- General search principles - not limited to sequence (or to biology).
- Start with broad keywords and narrow the search using more specific terms.
- Try variants of spelling, numbers, etc.
- Search many databases.
- Be persistent!!
- Use latest database version.
- Use BLAST first, then a finer tool (FASTA,...)
- Search both strands when using FASTA.
- Translate sequences where relevant
- Search 6-frame translation of DNA database
- $E < 0.05$  is statistically significant, usually biologically interesting.
- If the query has repeated segments, delete them and repeat search