Macromolecular and Physical Data

Michael J. Watts

http://mike.watts.net.nz

Lecture Outline

- Basic biochemistry
- Sources of biochemical data
- Representation of biochemical data
- Uses of biochemical data

What is DNA?

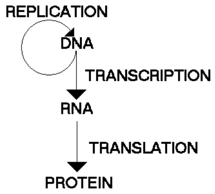
- DeoxyriboNucleicAcid
- storage medium of genetic information in higher organisms
- encapsulated in cell nucleus of eukaryotic (multi-cellular) organisms
- consists of long chains of nucleotides
- Two chains in double helix structure

What is DNA? (continued)

- Four bases:
  - •Adenine A
  - •Guanine G
  - •Thymine T
  - •Cytosine C
- sequence of bases encodes genetic information

How is DNA Processed in Cells?

- “Central Dogma of Molecular Genetics”
- describes the flow of information from DNA to protein


REPLICATION
DNA
TRANSCRIPTION
RNA
TRANSLATION
PROTEIN

DNA Processing (continued)

- DNA is transcribed into messenger RNA (mRNA)
  - •RNA is a less stable relative of DNA
  - •replaces Thymine (T) with Uracil (U)
- RNA strand read by ribosome to produce protein (translation)

Transcription

- DNA split into single strands
- RNA polymerase binds to DNA strand at promoter site
- RNA strand formed from DNA base complements
  - •A -> U G -> C
  - •C -> G T -> A

Transcription (continued)

- mRNA cut into sections
- Coding (exon) portions of RNA strands spliced together
- Non-coding (intron) segments discarded

RNA Translation

- Triplets of RNA bases (codons) translated to amino acid (residue)
  - •The genetic code
  - •Amino acids linked to form protein
- Protein folds according to electrostatic forces
- Shape of protein determines its function

DNA Data

- Many different kinds of DNA data and DNA related data in existence
- DNA promoter data
- RNA splice junction data
- The “genetic code”
- Protein sequences and configurations

DNA Data

- Lists of bases

```
AAGCTTCGTGAGCTGCGTAGGCTAGGGCTTTAGGCTCCGAGTC
CGTAAGCTCGAGACTGCTAGAGCTCTAGAGCTATAGCGCTATAC
GGACTATCGAGCTCTGGGCTATATATTTTATCGCGTTATAGAGA
GATCTCGAGATCGCGCGATCGAGCTTAGCAGCTATATCGGCTAT
CAGGCATCATAGCTTCGTGAGCTGCGTAGGCTAGGGCTTTAGG
CTCCGAGTCCGTAAGCTCGAGACTGCTAGAGCTCTAGAGCTATA
GCGCTATACGGACTATCGAGCTCTGGGCTATATATTTTATCGCG
TTATAGAGAGATCTCGAGATCGCGCGATCGAGCTTAGCAGCTAT
ATCGGCTATCAGGCATCAT
```

Sources of DNA Data

- EMBL
  - http://www.ebi.ac.uk/embl/Access/index.html
- BLAST
  - http://www.ncbi.nlm.nih.gov/BLAST/
- GenBank
  - http://www.ncbi.nlm.nih.gov/Genbank/

Uses of DNA Data

- gene finding
- disease detection
- disease prediction
- genetic engineering
- gene therapy?

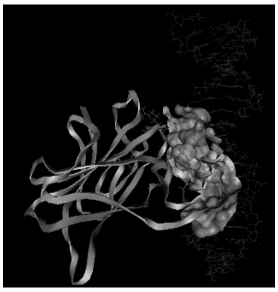
Protein Data

- Long chains of amino acids (residues)
- Can be sequenced
  - yields long lists of residues

ALA VAL SER LYS VAL TYR ALA ARG SER VAL TYR ASP SER
ARG GLY ASN PRO THR VAL GLU VAL GLU LEU THR THR GLU
LYS GLY VAL PHE ARG SER ILE VAL PRO SER GLY ALA SER
THR GLY VAL HIS GLU ALA LEU GLU MET ARG ASP GLY ASP
LYS SER LYS TRP MET GLY LYS GLY VAL LEU

Protein Data

- Proteins have complex 3D shapes



3D shape of protein (conformation) affects its biological function

Sources of Protein Data

- PDB
  - Protein Data Bank, Brookhaven
  - http://www.rcsb.org/pdb/
- SwissProt
  - http://au.expasy.org/sprot/
- Proteome, Inc.
  - http://www.proteome.com/

Uses of Protein Data

- Proteins are the means by which genes are expressed
- Study of proteins tells us how genes affect the organism
- Proteomics
  - MUCH larger field than genomics

Representing Biological Data

- Depends what you’re doing with it
- Data can be processed by statistical methods
  - need some numeric representation
- Data can be visualised
  - need to represent base identity

Representing Biological Data

- Basic sequence data string of letters
  - •A, C, G, T (DNA)
  - •A,C,D,E, etc for Amino Acids
- Can be represented in several ways
- Substitute arbitrary numbers for letters
  - •e.g. A=1, C=2, G=3, T=4
  - •doesn’t reflect some properties of the bases
  - •problems dealing with uncertainty

Representing Biological Data

- Binary representation
  - •orthogonal encoding
  - •each base can be one of four (or 20)
  - •represent each base by four bits
    - •i.e. A = 1 0 0 0, C = 0 1 0 0, etc.
  - •handles uncertainty better
    - •e.g. A or C = 1 1 0 0
  - •still ignores properties of the bases

Representing Biological Data

- Electron Ion Interaction Potential (EIIP)
  - •Measure of the chemical properties of DNA bases
  - •Preserves information about the properties of the bases
  - •specific to DNA
  - •Problems with uncertainty remain

Representing Biological Data

- Charge, hydrophobicities
  - – biophysical properties of amino acids
  - – specific to proteins

Representing Biological Data

- Most biochem / molecular biology databases are badly organised
- Use flat text files to store data
- Poor search facilities
- Major legacy problems
- Big opportunity for INFO graduates

Processing Biological Data

- Homology finding
- Gene expression
- Protein structure
- Gene mutations

Homology

- Finding similarity between sequences
- Implies evolutionary relationships between species
- Done via sequence alignment
- Sequence alignment compares two or more sequences

Sequence Alignment

- Goal is to maximise number of matching characters in each sequence
- Count number of matches and differences
- Differences include
  - substitutions
  - additions
  - Deletions

Sequence Alignment

- Substitution
  - a character has been replaced with another
- Addition (or insertion)
  - a character has been inserted into the sequence
- Deletion
  - a character has been removed from the sequence

Sequence Alignment

- This makes sequence alignment a difficult problem
- The sequences may be of different lengths
- Need to identify and align homologous groups / modules
  - local vs global alignment

Gene Expression

- Each cell nucleus contains a complete copy of the organism's genome
- BUT only specific genes are expressed in each tissue / cell type
- Gene expression analysis seeks to identify these genes

Gene Expression

- What use is this?
- Gene expression in tumors
  - find genes that are expressed in cancers and not other tissues
  - target drugs to these genes
  - disable tumors

Protein Structure

- Shape of protein determines its function
- Enzyme activity
  - enzymes bind to substrates
  - lock and key arrangement
  - shapes must match
  - therapeutic applications

Protein Structure

- Structure can be determined directly
  - crystallise the protein
  - examine with X-ray diffraction or NMR
- Difficult
  - not all proteins crystallise well
- Slow!

Protein Structure

- 'Holy Grail' of proteomics
  - predicting structure accurately from primary sequence
- Homology / alignments used so far
- Intelligent techniques can also be used
  - EC
  - ANN

Gene Mutations

- Single Nucleotide Polymorphisms
- Point mutations
- Change of a single nucleotide
- Examples
  - Sickle-cell disease
  - Lactose tolerance

Gene Mutations

- Can be used to diagnose disease
  - Sickle-cell disease
  - Huntington's chorea
- Can be used to predict disease
  - Stomach cancer
  - Alzheimers

Gene Mutations

- Targets for gene therapy
- Speculative treatment
- Uses retroviruses to replace defective genes
- No major successes yet
- Use of nanotechnology has been proposed..

Gene Mutations

- Can be discovered via sequence alignment
- Can be discovered via analysis of gene expression
- Once discovered, can be screened for (relatively) easily

Summary

- Bioinformatics is a very large field
- Filled with many challenges
  - and lots of $$$$!
- HUGE amount of data exists and being continuously produced
- Problems with processing it all
- Many opportunities for INFO-type folk