Intelligent Systems for Bioinformatics

Michael J. Watts

http://mike.watts.net.nz

Lecture Outline

- What is bioinformatics?
- What is DNA?
- How is it processed in cells?
- What is DNA data?
- How is DNA data represented?
- How can IS be applied to DNA data?

What is Bioinformatics?

- Computational analysis of biological sequences
- Many different kinds of biological data exist
- Amount of data increasing at an exponential rate
- Need automated methods of processing it

What is DNA?

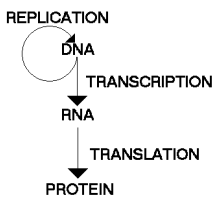
- DeoxyriboNucleicAcid
- storage medium of genetic information in higher organisms
- encapsulated in cell nucleus of eukaryotic (multi-cellular) organisms
- consists of long chains of nucleotides
- Two chains in double helix structure

What is DNA?

- 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



DNA Processing

- 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

- 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 it's 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

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
  - •Theoretical problems (measurement theory)

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
  - been tried in the past, but never really successful
  - specific to proteins

Applications of IS to Biological Data

- IS can be applied to each stage of the protein synthesis process
- identifying DNA promoter sites
- identifying RNA splice sites
- modelling the genetic code
- predicting protein configuration

Identifying DNA Promoter Sites

- Promoters are the start of coding regions of DNA
- DNA coding regions have known termination points
  - •typically AATAAA
- ANN can be trained to classify a region of DNA as promoter or non-promoter

RNA Splice Site Prediction

- Splicing is the second step in removing unused information
- Sites can be either intron - exon (non-coding - coding) or
- exon - intron (coding - non-coding)
- MLPs and Knowledge based neural networks (KBNN) have both been applied

RNA Splice Site Prediction

- Data set consists of sequences of 60 nucleotides
- each sequence represents either
  - •intron - exon junction
  - •exon - intro junction
  - •non-junction region
- binary representation of bases used
- Performance exceeds statistical methods

Modelling the Genetic Code

- 3 bases in each codon (64 combinations)
- 20 natural amino acids plus STOP codon
  - •genetic code is degenerate
- Train a MLP to model the genetic code
- Problems with training indicate properties of the genetic code
- Internal representation matches known biochemical groupings

Protein Configuration Prediction

- Proteins are formed from chains of amino acids
- Proteins have primary, secondary, tertiary and quaternary structures
- Primary structure is its amino acid sequence
- Electrostatic forces folds it into secondary, tertiary or even quaternary structures

Protein Configuration Prediction

- Final structure determines it's biological function
- Secondary structure consists of sub structures
  - •alpha helix
  - •pleated (or beta) sheet
  - •rest is coil
- Can use ANN to predict the structure from amino acid sequence

Secondary Structure Prediction

- Modelled with an ANN in 1988
- Used an MLP to predict secondary structure classes
- Input was a 'window' of 13 residues
- Predict which SS centre of the window is in

Secondary Structure Prediction

- Each residue represented by a 21-bit vector
  - 1 bit for each residue type
  - +1 for 'spacer' bit
- Total input size of 273 bits
- 3 output nodes
  - helix, sheet, coil

Secondary Structure Prediction

- Sequence profiles
  - Find a set of known proteins that are similar (homologous) to the unknown protein
  - Align the homologues with the unknown so that the maximum number of residues match
  - Use the alignment to construct a profile of residue frequencies

Secondary Structure Prediction

- Use the profile and other measures as inputs to the ANN
- This gives the best prediction results so far (>86% accuracy)
- Used in the PhD prediction server
  - e-mail based prediction server

Protein Configuration Prediction

- Genetic algorithms used to predict tertiary structure
- fitness based on the energy required to maintain each structure
  - •GA aims to minimise energy
- Alternative is to use brute force search
  - •*very* time consuming

Signal Peptide Cleavage

- Proteins that are exported from a cell are "marked"
  - "signal" molecule at end of protein
  - cleaved off before export
- ANN can be used to predict where the "signal" section ends
- Organism specific, to some extent

Signal Peptide Cleavage

- Prediction is based on structure, rather than sequence
- Statistical methods have problems modelling this kind of thing
- ANN have shown some promise
  - Still work to be done

Conclusion

- 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
- Intelligent systems can be used to do this