Bioinformatics 2

Introduction

Lecture 1

- Course Overview & Assessment
- Introduction to Bioinformatics
- Careers and PhD options
- Introduction to Sequence Alignment

About me...

- Started in Biology (behaviour genetics)
- Got interested in databases (anatomy)
- Commercial and Academic Experience
- 'wet lab' and bioinformatics projects
- Office in FH, Lab in HRB

The class (2006)

- M.Sc. Classes:
- Quantitative Genetics and Genome Analysis (assignment 1 and 2)
- Bioinformatics 2 (assignment 1 and exam)

What do I think you know?

- Variety of backgrounds and experience:
  - Biological Sciences
  - Computing Sciences
  - Mathematics, Statistics and Physics

Course Outcomes

- Know the core algorithms in bioinformatics
- Experience in using and/or implementing simple solutions
- Appreciate the current 'state of the art'
  - what has been solved?
  - what are the key limitations?
- Be familiar with the available resources

Course Design

- Lectures cover essential background
- Guest lectures present research level
- Self-study and assignments designed to cover practical implementation

Assessment (Bio2)

- Written exam 70%
- Written assignment 30%
  - Experimental design and data analysis mini project
- Plagiarism will be refereed externally

Late submissions will be penalized

Bioinformatics?

- What is Bioinformatics?
- What does Bioinformatics do for CS?
- What does Bioinformatics do for Biology?
- What guest lecture would you like?

- Discuss in groups for 15 min.

What is BioInformatics?

- Sequence analysis and genome building
- Molecular Structure prediction
- Evolution, phylogeny and linkage
- Automated data collection and analysis
- Simulations
- Biological databases and resources

BioInf and CS

- Provides CS with new challenges with clear medical significance.
- Complex and large datasets sometimes very noisy with hidden structures.
- Can biological solutions be used to develop new computational tools?

BioInf and Biology

- High-throughput biology:
  - around 1989, the sequence of a 1.8kb gene would be a PhD project
  - by 1993, the same project was an undergraduate project
  - in 2000 we generated 40kb sequence per week in a non-genomics lab.

BioInf and Biology

- High-throughput biology
- Data management and mining
- Modeling of Biological theories
- Analysis of complex systems

Bioinformatics@ed

- Database integration
- Data provenance
- Evolutionary and genetic computation
- Gene expression databases
- High performance data structures for semi-structured data (Vectorised XML)

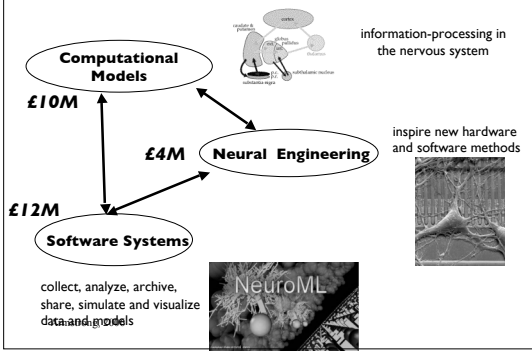
1/2

Bioinformatics@ed

- Machine learning
- Microarray data analysis
- Natural language and bio-text mining
- Neural computation, visualisation and simulation
- Protein complex modeling

2/2

Neuroinformatics



Career Options

- Academic Routes
  - Get Ph.D, do Postdoctoral Research - lectureship and independent group
  - M.Sc. RA - becomes semi independent usually linked to one or more academic groups. Career structure is less defined but improving. RAs can do Ph.D. part-time.

Career Options

- Commercial Sector
  - Big Pharma - Accept PhD and MSc entry. Normally assigned to projects and work within defined teams. Defined career structure (group leaders, project managers etc)
  - Spin-out/Small biotech - Accept PhD and MSc entry. More freedom and variety. A degree of 'maintenance' work is to be expected.

Career Options

- Hybrid Approaches
  - Commercial and Academic research groups are becoming much closer linked.
  - University academics encouraged to exploit their IPR (intellectual property rights).
  - Companies can get government support to collaborate with academic research groups.

Ph.D.

- Assuming a start date of October 2006
- 'prize' studentships advertised on jobs.ac.uk, nature, science etc starting NOW!
  - Many linked to nationality/residency (Check details carefully).
- UK 'quota' studentships vary with department but contact/apply early.

Ph.D.

- US studentships take longer but are better paid and have extra training/coursework
  - require an entry exam
  - again, deadlines are very soon for '06

Sequence Alignment

What is it?

```
ACCGGTATCCTAGGAC
ACCTATCTTAGGAC
```

Are these two sequences related?
How similar (or dissimilar) are they?

What is it?

```
ACCGGTATCCTAGGAC
|||  ||||  ||||||
ACC--TATCTTAGGAC
```

- Match the two sequences as closely as possible = aligned
- Therefore, alignments need a score

Why do we care?

- DNA and Proteins are based on linear sequences
- Information is encoded in these sequences
- All bioinformatics at some level comes back to matching sequences that might have some noise or variability

Alignment Types

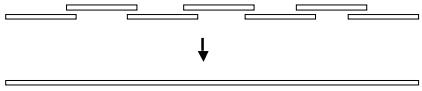
- Global: used to compare to similar sized sequences.
  - Compare closely related genes
  - Search for mutations or polymorphisms in a sequence compared to a reference.

Alignment Types

- Local: used to find shared subsequences.
  - Search for protein domains
  - Find gene regulatory elements
  - Locate a similar gene in a genome sequence.

Alignment Types

- Ends Free: used to find joins/overlaps.
  - Align the sequences from adjacent sequencing primers.



How do we score alignments?

ACCGG**TATC**C**TAGGAC**

||| |||| ||||||

ACC--TATCT**TAGGAC**

- Assign a score for each match along the sequence.

How do we score alignments?

ACCGGTATC**C**TAGGAC

||| |||| ||||||

ACC--TATC**T**TAGGAC

- Assign a score (or penalty) for each substitution.

How do we score alignments?

```
ACCGGTATCCTAGGAC
|||  ||||  ||||||
ACC--TATCTTAGGAC
```

- Assign a score (or penalty) for each insertion or deletion.
- insertions/deletions otherwise known as indels

How do we score alignments?

```
ACCGGTATCCTAGGAC
|||  ||||  ||||||
ACC--TATCTTAGGAC
```

- Matches and substitutions are 'easy' to deal with.
  - We'll look at substitution matrices later.
- How do we score indels: gaps?

How do we score gaps?

```
ACCGGTATCC---GAC
|||  ||||    |||
ACC--TATCTTAGGAC
```

- A gap is a consecutive run of indels
- The gap length is the number of indels.
- The simple example here has two gaps of length 2 and 3

How do we score gaps?

```
ACCGGTATCC---GAC
|||  ||||    |||
ACC--TATCTTAGGAC
```

- Constant: Length independent weight
- Affine: *Open* and *Extend* weights.
- Convex: Each additional gap contributes less
- Arbitrary: Some arbitrary function on length

Choosing Gap Penalties

- The choice of Gap Scoring Penalty is very sensitive to the context in which it is applied:
  - introns vs exons
  - protein coding regions
  - mis-matches in PCR primers

Substitution Matrices

- Substitution matrices are used to score substitution events in alignments.
- Particularly important in Protein sequence alignments but relevant to DNA sequences as well.
- Each scoring matrix represents a particular theory of evolution

Similarity/Distance

- Distance is a measure of the cost or replacing one residue with another.
- Similarity is a measure of how similar a replacement is.

e.g. replacing a hydrophobic residue with a hydrophilic one.

- The logic behind both are the same and the scoring matrices are interchangeable.

DNA Matrices

Identity matrix

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

BLAST

	A	C	G	T
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5

However, some changes are more likely to occur than others (even in DNA). When looking at distance, the ease of mutation is a factor. a.g. A-T and A-C replacements are rarer than A-G or C-T.

Frameshifts

DNA frameshifts

Jean-Michael Claverie (1993). Detecting frame shifts by amino acid sequence comparison. J Mol Biol 234 1140-1157

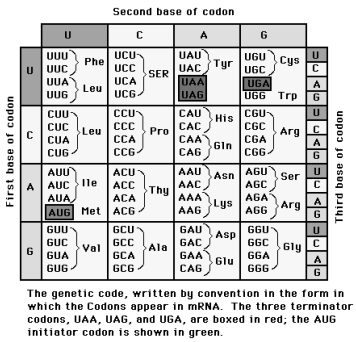
Created matrices for detecting frame shift mutations that give rise to new coding sequences or that arise from sequencing errors.

Protein Substitution Matrices

How can we score a substitution in an aligned sequence?

- Identity matrix like the simple DNA one.
- Genetic Code Matrix:

  For this, the score is based upon the minimum number of DNA base changes required to convert one amino acid into the other.



The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

Protein Substitution Matrices

How can we score a substitution in an aligned sequence?

- Amino acid property matrix

  Assign arbitrary values to the relatedness of different amino acids:

  e.g. hydrophobicity , charge, pH, shape, size

Matrices based on Probability

$S_{ij} = \log (q_{ij}/p_i p_j)$

Sij is the log odds ratio of two probabilities: amino acids i and j are aligned by evolutionary descent and the probability that they aligned at random**.**

This is the basis for commonly used substitution matrices.

PAM matrices

Dayhoff, Schwarz and Orcutt 1978 took these into consideration when constructing the PAM matrices:

Took 71 protein families - where the sequences differed by no more than 15% of residues (i.e. 85% identical)

- Aligned these proteins
- Build a theoretical phylogenetic tree
- Predicted the most likely residues in the ancestral sequence

PAM Matrices

- Ignore evolutionary direction
- Obtained frequencies for residue X being substituted by residue Y over time period Z
- Based on 1572 residue changes
- They defined a substitution matrix as 1 PAM (point accepted mutation) if the expected number of substitutions was 1% of the sequence length.

PAM Matrices

To increase the distance, they multiplied the the PAM1 matrix.

PAM250 is one of the most commonly used.

PAM - notes

The PAM matrices are rooted in the original datasets used to create the theoretical trees

They work well with closely related sequences

Based on data where substitutions are most likely to occur from single base changes in codons.

PAM - notes

Biased towards conservative mutations in the DNA sequence (rather than amino acid substitutions) that have little effect on function/structure.

Replacement at any site in the sequence depends only on the amino acid at that site and the probability given by the table.This does not represent evolutionary processes correctly. Distantly related sequences usually have regions of high conservation (blocks).

PAM - notes

36 residue pairs were not observed in the dataset used to create the original PAM matrix

A new version of PAM was created in 1992 using 59190 substitutions: Jones, Taylor and Thornton 1992 CAMBIOS 8 pp 275

BLOSUM matrices

Henikoff and Henikoff 1991

Took sets of aligned ungapped regions from protein families from the BLOCKS database.

The BLOCKS database contain short protein sequences of high similarity clustered together. These are found by applying the MOTIF algorithm to the SWISS-PROT and other databases. The current release has 8656 Blocks.

BLOSUM matrices

Sequences were clustered whenever the %identify exceeded some percentage level.

Calculated the frequency of any two residues being aligned in one cluster also being aligned in another

Correcting for the size of each cluster.

BLOSUM matrices

Resulted in the fraction of observed substitutions between any two residues over all observed substitutions.

The resulting matrices are numbered inversely from the PAM matrices so the BLOSUM50 matrix was based on clusters of sequence over 50% identity, and BLOSUM62 where the clusters were at least 62% identical.

Summary so far…

- Gaps
  - Indel operations
  - Gap scoring methods
- Substitution matrices
  - DNA largely simple matrices
  - Protein matrices are based on probability
  - PAM and BLOSUM

Final Note

- We try to keep one lecture slot (minimum) free for guest lectures.
- If you have any subject you would like us to try and cover then let me know asap.
- Current possibilities are:
  - Commercial perspectives on Bioinformatics
  - Drug design and Bioinformatics