Introduction to bioinformatics

Graham J.L. Kemp

11 November 2013

What is bioinformatics?

"Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data."

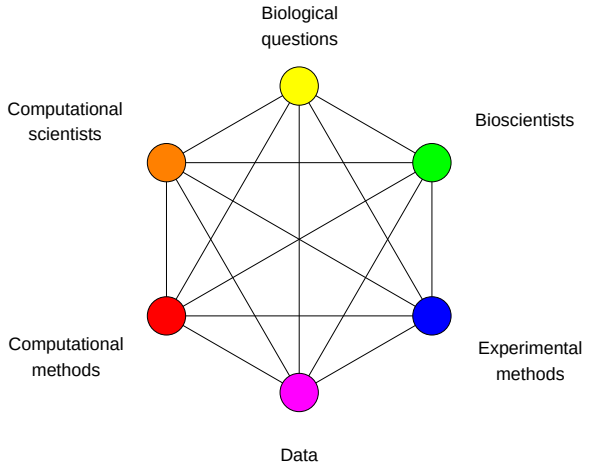
"Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful."

Working definition by the NIH Biomedical Information Science and Technology Initiative Consortium, 2000
http://www.bisti.nih.gov/docs/CompuBioDef.pdf

What is biology?

Ecosystem	Rain forest, desert, fresh water lake, digestive tract of an animal
Community	All species in an ecosystem
Population	All individuals of a single species
Organism	One single individual
Organ System	A specialised functional system of an organism, e.g. nervous system or immune system
Organ	A specialised structural system of an organism, e.g. brain or kidney
Tissue	A specialised substructure of an organ, e.g. nervous tissue, smooth muscle
Cell	A single cell, e.g. neuron, skin cell, stem cell, bacteria
Molecule	e.g. protein, DNA, RNA, sugar, fatty acid, metabolites, pharmaceutical drugs

Addressing biological questions



Sequences, structures and systems

Sequences

- Nucleic acids (DNA and RNA) and proteins are (unbranched) polymers. Their composition can be described by the sequence of units (nucleotides or amino acid residues) in a chain.

Structures

- Three-dimensional structures can give insights into the molecular basis of biological functions.

Systems

- Biological processes consist of the coordinated actions of molecules.

Biological sequences: some experimental methods

- DNA sequencing
- Protein sequencing
- Next-generation sequencing (NGS)

Biological sequences: some questions

- How similar are a pair of sequences?
- Identify the corresponding units in a pair of homologous molecules that have undergone substitutions and insertions/deletions during their evolutionary history (*pairwise sequence alignment*).
- Given a new sequence, has anything similar (in whole or part) been seen before?
- Reconstruct a phylogenetic tree from the sequences of a set of homologous molecules.
- Given the sequences of many overlapping DNA fragments from a single organism, assemble them to reconstruct a full genome.
- Given the sequences of many DNA fragments from a mixture of organisms, identify the species present in the mixture.

Biological structures: some experimental methods

Find the atomic structure of a macromolecule or complex

- X-ray crystallography
- Nuclear magnetic resonance (NMR) spectroscopy

Identify a low-resolution "envelope" enclosing a large macromolecular complex

- Cryo-electron microscopy
- Small-angle x-ray scattering

Biological structures: some questions

- Can differences in the functions of two similar proteins be explained by differences in their structures?
- Can a drug be designed to fit into the active site of a target protein?
- Can the safety and efficacy of a potential therapeutic protein be predicted from its structure?
- Can the function of a protein be altered by changing its composition, and hence its structure?
- Can a protein's structure be predicted from its sequence?
  - the protein folding problem
- Given the structures of two proteins, will they associate with one another? If so, how will they fit together?
  - the protein docking problem

Biological systems: some experimental methods

Which mRNA molecules are being expressed?

- Microarray gene expression
- RNA-Seq

Which proteins are being expressed?

- (2-D) gel electrophoresis
- Mass spectrometry

In which tissue(s) are particular genes expressed?

- *in situ* hybridization

Biological systems: some questions

- Which genes/proteins are co-expressed (i.e. have similar expression profiles)?
- Which genes are expressed in tumour cells but not in healthy cells?
- If a gene is "knocked out", will an organism survive, and how will the expression of other genes be affected?
- Can protein expression profiles identify proteins that could be targets for drug development?
- Can an individual's expression profile indicate whether they are likely to respond to a particular therapeutic treatment?
- How do biological networks respond to injury or to treatment with a therapeutic drug?

Bioinformatics in TMS145

Focus on two important kinds of biological data:

- biological sequences
- macromolecular structures

Using data to understand evolutionary relationships

- Lab on sequence alignment; comparing related biological sequence
- Lab on protein structure comparison; examining similarities and differences in a family of related protein stuctures

Sequence alignment

Comparison of macromolecular sequences.

Nucleic acids (DNA, RNA) or proteins.

Assignment of nucleotide-nucleotide or residue-residue correspondences.

Suggest evolutionary, structural and functional relationships.

Rigorous algorithms for global and local alignment.

Heuristic algorithms for practical database searching.

Dotplots

A pictorial representation of the similarity between two sequences.

Compare a sequence with itself:
- Repeats
- Palindromic sequences

Compare two sequences:
- Any path from upper left to lower right represents an alignment.
- Horizontal or vertical moves correspond to gaps in one of the sequences.
- Path with highest score corresponds to an optimal alignment.

Measures of sequence similarity

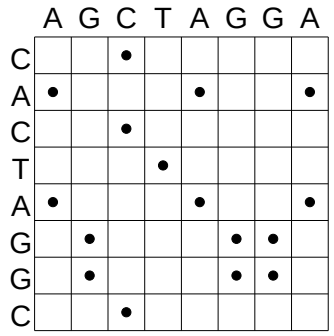
Hamming distance:
- Number of positions with mismatching characters.
- Defined for two strings of equal length.

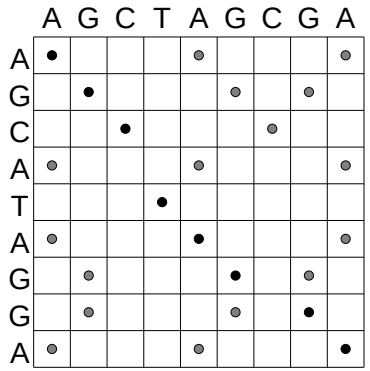
```
agtc
cgta
```

Levenshtein distance:
- Minimum number of edit operations (delete, insert, change a single character) needed to change one sequence into another.

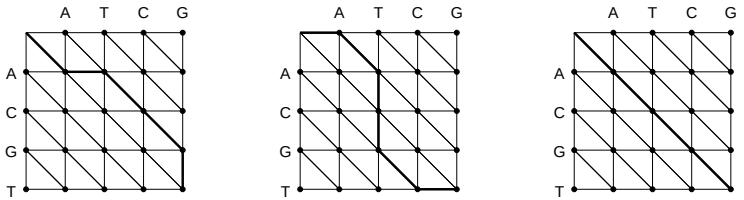
```
agtcc
cgctca
```


Dotplots





Each path represents an alignment




```
A-CGT
| ||
ATCG-
```



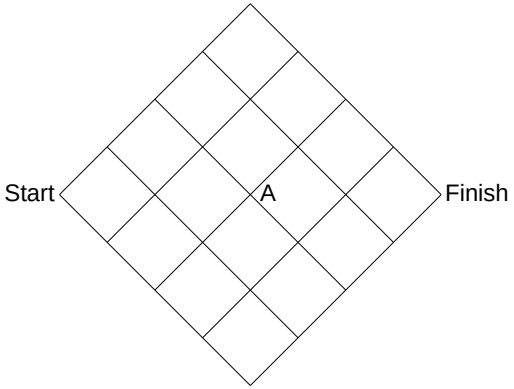
```
-ACGT-

AT--CG
```



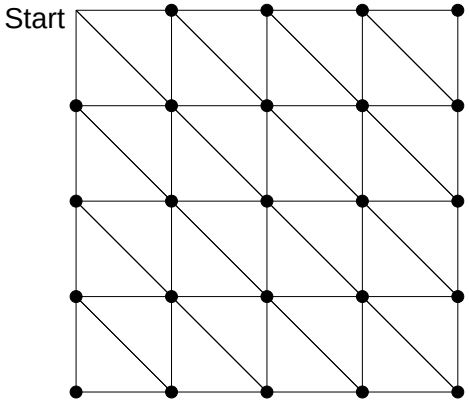
```
ACGT
|
ATCG
```

- Vertical steps add a gap to the horizontal sequence
- Horizontal steps add a gap to the vertical sequence

Do we have to enumerate all paths?



How many paths?



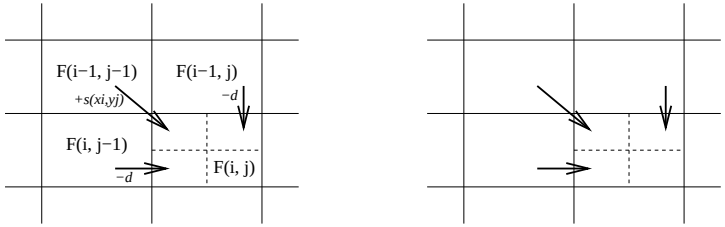
Pairwise global alignment (Needleman-Wunsch algorithm)

Rigorous algorithms use dynamic programming to find an optimal alignment.

- match score
- mismatch score
- gap penalty

$$F(i,\,j) = \max \begin{cases} F(i-1,\,j-1) + s(x_i, y_j) \\ F(i-1,\,j) - d \\ F(i,\,j-1) - d \end{cases}$$

Dynamic programming



Percent identity

Having obtained an alignment, it is common to quantify the similarity between a pair of sequences by stating the percent identity.

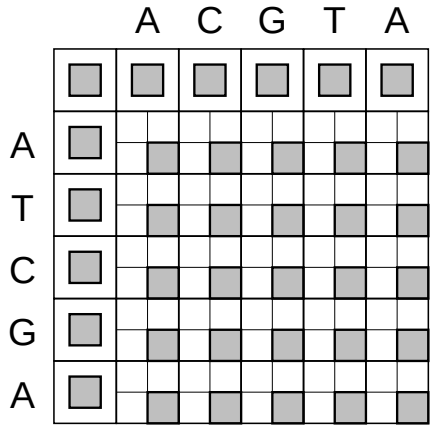
Count the number of alignment positions with matching characters and divide by ... *what*?

- the length of the shortest sequence?
- the length of the alignment?
- the average length of the sequences?
- the number of non-gap positions?
- the number of equivalenced positions excluding overhangs?

Sequences are either homologous (i.e. they share a common evolutionary ancestor) or they are not.

The phrase “percent homology” is meaningless!

Score matrix



Pairwise local alignment (Smith-Waterman algorithm)

Local similarities may be masked by long unrelated regions.

A minor modification to the global alignment algorithm.

- If the score for a subalignment becomes negative, set the score to zero.

$$F(i,\,j) = \max \begin{cases} 0 \\ F(i-1,\,j-1) + s(x_i, y_j) \\ F(i-1,\,j) - d \\ F(i,\,j-1) - d \end{cases}$$

- Trace back from the position in the score matrix with the highest value.
- Stop at cell where score is zero.