Medical Informatics

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

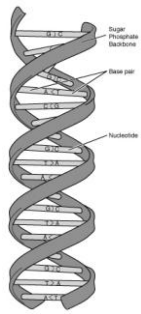
http://www.yildiz.edu.tr/~naydin

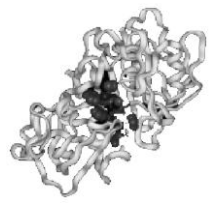
1

BIOINFORMATICS

2

What is Bioinformatics?...





3

…What is Bioinformatics?...

- Bioinformatics
  - the study of how information is represented and transmitted in biological systems, starting at the molecular level

  is a discipline that does not need a computer.
  - An ink pen and a supply of traditional laboratory notebooks could be used to record results of experiments.
  - However, to do so would be like foregoing the use of a computer and word-processing program in favor of pen and paper to write a novel.

4

…What is Bioinformatics?...

- From a practical sense, bioinformatics is a science that involves
  - collecting,
  - manipulating,
  - analyzing,
  - transmitting

  huge quantities of data,
- uses computers whenever appropriate.
- bioinformatics refers to computational bioinformatics.

5

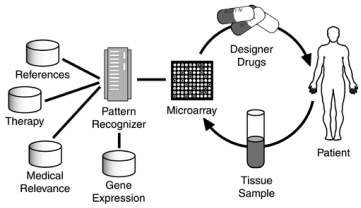
Killer application

- In the biotechnology industry, every researcher and entrepreneur hopes to develop or discover the next "killer app“
  - the one application that will bring the world to his or her door and provide funding for R&D, marketing, and production.
    - For example, in general computing, the electronic spreadsheet and the desktop laser printer have been the notable killer apps.
    - The spreadsheet not only transformed the work of accountants, research scientists, and statisticians, but the underlying tools formed the basis for visualization and mathematical modeling.
    - The affordable desktop laser printer created an industry and elevated the standards of scientific communications, replacing rough graphs created on dot-matrix printers with high-resolution images.

6

Killer application

- "What might be the computer-enabled 'killer app' in bioinformatics?"
- Although there are numerous military and agricultural opportunities, one of the most commonly cited examples of the killer app is in personalized medicine, as illustrated in Figure



personalized medicine: the custom, just-in-time delivery of medications (popularly called "designer drugs") tailored to the patient'scondition.

7

Killer application

- Instead of taking a generic or over-the-counter drug for a particular condition,
  - a patient would submit a tissue sample, such as a mouth scraping, and submit it for analysis.
    - A microarray would then be used to analyze the patient's genome and the appropriate compounds would be prescribed.
- The drug could be a cocktail of existing compounds, much like the drug cocktails used to treat cancer patients today.
- Alternatively, the drug could be synthesized for the patient's specific genetic markers—
  - as in tumor specific chemotherapy, for example.
    - This synthesized drug might take a day or two to develop, unlike the virtually instantaneous drug cocktail.
    - The tradeoff is that the drug would be tailored to the patient's genetic profile and condition, resulting in maximum response to the drug, with few or no side effects.

8

Killer application

- How will this or any other killer app be realized?
  - The answer lies in addressing the molecular biology, computational, and practical business aspects of proposed developments such as custom medications.
- A practical system would include:
  - High throughput screening
    - The use of affordable, computer-enabled microarray technology to determine the patient's genetic profile.
    - The issue here is affordability, in that microarrays costs tens of thousands of dollars

9

Killer application

  - Medically relevant information gathering
    - Databases on gene expression, medical relevance of signs and symptoms, optimum therapy for given diseases, and references for the patient and clinician must be readily available.
    - The goal is to be able to quickly and automatically match a patient's genetic profile, predisposition for specific diseases, and current condition with the efficacy and potential side effects of specific drug-therapy options.
  - Custom drug synthesis
    - The just-in-time synthesis of patient-specific drugs, based on the patient's medical condition and genetic profile, presents major technical as well as political, social, and legal hurdles.
    - For example, for just-in-time synthesis to be accepted by the FDA, the pharmaceutical industry must demonstrate that custom drugs can skip the clinical-trials gauntlet before approval.

10

Killer application

- Achieving this killer app in biotech is highly dependent on
  - computer technology,
    - especially in the use of computers to speed the process testing-analysis-drug synthesis cycle, where time really is money.
- For example, consider that for every 5,000 compounds evaluated annually by the U.S. pharmaceutical R&D laboratories, 5 make it to human testing, and only 1 of the compounds makes it to market.

11

Killer application

- In addition, the average time to market for a drug is over 12 years,
  - including several years of pre-clinical trials followed by a 4-phase clinical trial.
- These clinical trials progress from
  - safety and dosage studies in Phase I,
  - to effectiveness and side effects in Phase II,
  - to long-term surveillance in Phase IV,

with each phase typically lasting several years.

12

Killer application

- Most pharmaceutical companies view computerization as the solution to creating smaller runs of drugs focused on custom production.
- Obvious computing applications range from
  - predicting efficacy and side effects of drugs based on genome analysis,
  - to visualizing protein structures to better understand and predict the efficacy of specific drugs,
  - to illustrating the relative efficacy of competing drugs in terms of quality of life and cost, based on the Markov simulation of likely outcomes during Phase IV clinical trials.

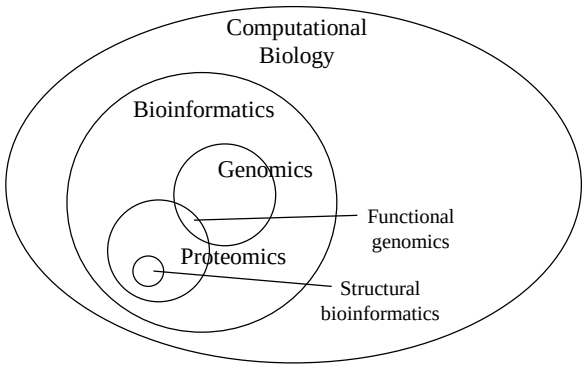
13

…What is Bioinformatics?...

- A quick google search with the keywords *bioinformatics*
  - yields about **1.480.000** results (6th October 2008)
  - yields about **24.900.000** results (7th February 2016)
- **Synonyms:**
  - Computational Biology
  - Computational Molecular Biology
  - Biocomputing

14

...What is Bioinformatics?...



15

...What is Bioinformatics?...

- Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information
  - stored in the genetic code,
  - experimental results from various sources,
  - patient statistics,
  - and scientific literature.
- Research in bioinformatics includes method development for
  - storage,
  - retrieval,
  - analysis

  of the data.

16

...What is Bioinformatics?...

- Bioinformatics
  - a rapidly developing branch of biology
  - highly interdisciplinary,
  - using techniques and concepts from
    - informatics,
    - statistics,
    - mathematics,
    - chemistry,
    - biochemistry,
    - physics,
    - linguistics.

17

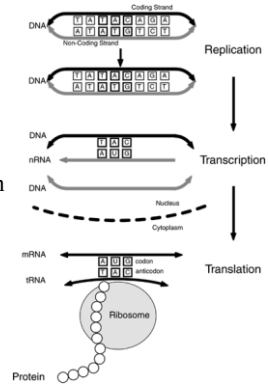
…What is Bioinformatics?...

- The relationship between computer science and biology is a natural one for several reasons.
  - 1st,
    - the phenomenal rate of biological data being produced provides challenges:
      - massive amounts of data have to be stored, analysed, and made accessible.
  - 2nd,
    - the nature of the data is often such that a statistical method, and hence computation, is necessary.
      - This applies in particular to the information on the building plans of proteins and of the temporal and spatial organisation of their expression in the cell encoded by the DNA.
  - 3rd,
    - there is a strong analogy between the DNA sequence and a computer program
      - it can be shown that the DNA represents a Turing Machine.

18

The Central Dogma of Molecular Biology

- DNA is transcribed to messenger RNA in the cell nucleus, which is in turn translated to protein in the cytoplasm.
- The Central Dogma, shown here from a structural perspective, can also be depicted from an information flow perspective



19

…What is Bioinformatics?...

- The National Center for Biotechnology Information (NCBI 2001) defines bioinformatics as:
  - "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline.
  - There are three important sub-disciplines within bioinformatics:
    - the development of new algorithms and statistics with which to assess relationships among members of large data sets;
    - the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures;
    - the development and implementation of tools that enable efficient access and management of different types of information."

20

…What is Bioinformatics?...

- **From Webopedia:**
  - The application of computer technology to the management of biological information.
  - Specifically, it is the science of developing computer databases and algorithms to facilitate and expedite biological research.
  - Bioinformatics is being used largely in the field of human genome research by the Human Genome Project that has been determining the sequence of the entire human genome (about 3 billion base pairs) and is essential in using genomic information to understand diseases.
  - It is also used largely for the identification of new molecular targets for drug discovery.

21

…What is Bioinformatics?...

- The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.
- *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.
- *Computational Biology:*
  - The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

22

…What is Bioinformatics?...

- **Path to the Bioinformatics**
  - 1st,
    - Learn Biology.
  - 2nd,
    - Decide and pick a problem that interests you for experiment.
  - 3rd,
    - Find and learn about the Bioinformatics tools.
  - 4th,
    - Learn the Computer Programming Languages.
      - Perl, Pyton, R, Java, etc.
  - 5th,
    - Experiment on your computer and learn different programming techniques.

23

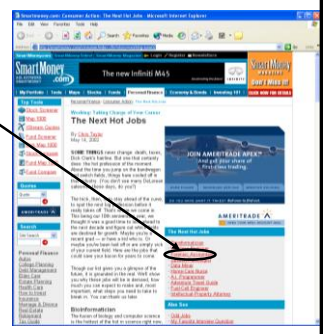
Why is Bioinformatics Important?

- Applications areas include
  - Medicine
  - Pharmaceutical drug design
  - Toxicology
  - Molecular evolution
  - Biosensors
  - Biomaterials
  - Biological computing models
  - DNA computing

24

Why should I care?

- SmartMoney ranks Bioinformatics as #1 among next HotJobs
- Business Week 50 Masters of Innovation
- Jobs available, exciting research potential
- Important information waiting to be decoded!



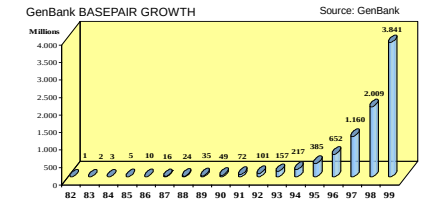
25

Why is bioinformatics hot?

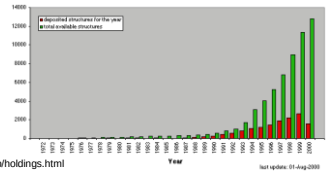
- Supply/demand: few people adequately trained in both biology and computer science
- Genome sequencing, microarrays, etc lead to large amounts of data to be analyzed
- Leads to important discoveries
- Saves time and money

26

The Role of *Computational* Biology



3D Structures Growth:



Source: http://www.rcsb.org/pdb/holdings.html

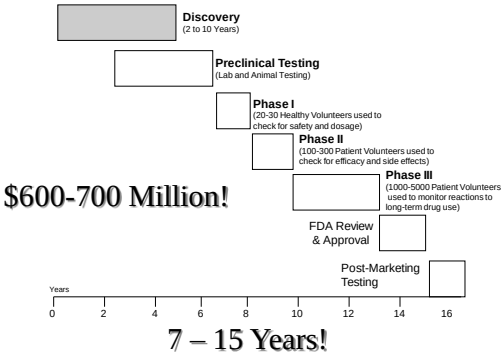
27

Fighting Human Disease

- Genetic / Inherited
  - Diabetes
- Viral
  - Flu, common cold
- Bacterial
  - Meningitis, Strep throat

28

Drug Development Life Cycle



$600-700 Million!

7 – 15 Years!

29

Drug lead screening

5,000 to 10,000 compounds screened

250 Lead Candidates in Preclinical Testing

5 Drug Candidates enter Clinical Testing; 80% Pass Phase I

30%Pass Phase II

80% Pass Phase III

One drug approved by the FDA

30

What skills are needed?

- Well-grounded in one of the following areas:
  - Computer science
  - Molecular biology
  - Statistics
- Working knowledge and appreciation in the others!

31

Where Can I Learn More?

- ISCB: http://www.iscb.org/
- NBCI: http://ncbi.nlm.nih.gov/
- http://www.bioinformatics.org/
- Journals
- Conferences

32

Some computational methods

- Introduction to sequence alignment
- pair wise sequence alignment
  - The Dot Matrix
  - Scoring Matrices
  - Gap Penalties
  - Dynamic Programming

33

Introduction to sequence alignment

Sequence Alignment is the identification of residue-residue correspondences.

- It is the basic tool of bioinformatics.

34

Sequence Alignment

- Question: Are two sequences related?
- Compare the two sequences, see if they are similar
- Example: ***pear*** and ***tear***
- Similar words, different meanings

35

Protein Evolution

"For many protein sequences, evolutionary history can be traced back 1-2 billion years"

-William Pearson

- When we align sequences, we assume that they share a common ancestor
  - They are then homologous
- Protein fold is much more conserved than protein sequence
- DNA sequences tend to be less informative than protein sequences

36

Use Protein Sequences for Similarity Searches

1) 4 DNA bases vs. 20 amino acids - less chance similarity
2) Similarity of AAs can be scored
   - # of mutations, chemical similarity, PAM matrix
3) Protein databanks are much smaller than DNA databanks
   -less random matches.
4) Similarity is determined by pairwise alignment of different sequences

37

Pairwise Alignment

- The alignment of two sequences (DNA or protein) is a relatively straightforward computational problem.
  - There are lots of possible alignments.
- Two sequences can always be aligned.
- Sequence alignments have to be scored.
- Often there is more than one solution with the same score.

38

Biological Sequences

- Similar biological sequences tend to be related
- Information:
  - Functional
  - Structural
  - Evolutionary
- Common mistake:
  - **sequence similarity is not homology!**
- Homologous sequences: derived from a common ancestor

39

Relation of sequences

- **Homologs**: similar sequences in 2 different organisms derived from a common ancestor sequence.
- **Orthologs**: Similar sequences in 2 different organisms that have arisen due to a speciation event. Functionality Retained.
- **Paralogs**: Similar sequences within a single organism that have arisen due to a gene duplication event.
- **Xenologs**: similar sequences that have arisen out of horizontal transfer events (symbiosis, viruses, etc)

40

Relation of sequences

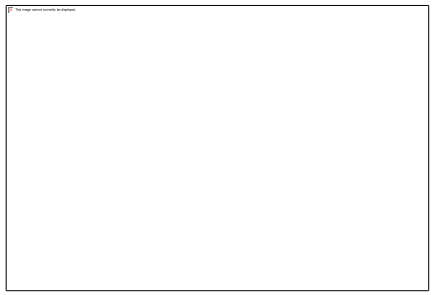


Image Source:
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

41

Sequence Alignment

The concept

- An alignment is a mutual arrangement of two sequences.
- It exhibits where the two sequences are similar, and where they differ.
- An 'optimal' alignment is one that exhibits the most correspondences, and the least differences.
- sequences that are similar probably have the same function

42

Sequence Alignment

Terms of sequence comparison

- **Sequence identity**
  - exactly the same Amino Acid or Nucleotide in the same position
- **Sequence similarity**
  - substitutions with similar chemical properties
- **Sequence homology**
  - general term that indicates evolutionary relatedness among sequences
  - sequences are homologous if they are derived from a common ancestral sequence
  - one speaks of percentage of sequence homology

43

Sequence Alignment

Things to consider:

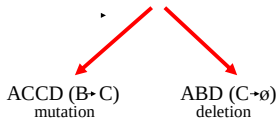
- to find the best alignment one needs to examine all possible alignments
- to reflect the quality of the possible alignments one needs to score them
- there can be different alignments with the same highest score
- variations in the scoring scheme may change the ranking of alignments

44

sequence alignment

Evolution:

Ancestral sequence: ABCD



ACCD
AB—D

or

ACCD
A—BD

Pairwise Alignment

true alignment

45

sequence alignment

A protein sequence alignment

```
MSTGAVLIY--TSILIKECHAMPAGNE-----
---GGILLFHRTHELIKESHAMANDEGGSNNS
   *  *    *  **** ***
```

A DNA sequence alignment

```
attcgttggcaaatcgcccctatccggccttaa
att---tggcggatcg-cctctacgggcc----
***   ****  **** **    ******
```

46

Edit Distance

- Sequence similarity: function of edit distance between two sequences

```
P E A R
  | | |
T E A R
```

47

Hamming Distance

- Minimum number of letters by which two words differ
- Calculated by summing number of mismatches
- Hamming Distance between PEAR and TEAR is 1

48

Gapped Alignments

- Biological sequences
  - Different lengths
  - Regions of insertions and deletions
- Notion of gaps (denoted by '-')

```
A L I G N M E N T
  | | |   | | | |
- L I G A M E N T
```

49

Possible Residue Alignments

- Match
- Mismatch (substitution or mutation)
- Insertion/Deletion (INDELS – gaps)

50

Alignments

- Which alignment is best?

```
A - C - G G - A C T
|   |   |       | |
A T C G G A T _ C T

A T C G G A T C T
|   | | |     | |
A - C G G - A C T
```

51

Alignment Scoring Scheme

- Possible scoring scheme:
  match: +2
  mismatch: -1
  indel –2
- Alignment 1: $5 * 2 – 1(1) – 4(2) = 10 – 1 – 8 = 1$
- Alignment 2: $6 * 2 – 1(1) – 2\ (2) = 12 – 1 – 4 = 7$

52

Alignment Methods

- Visual
- Brute Force
- Dynamic Programming
- Word-Based (k tuple)

53

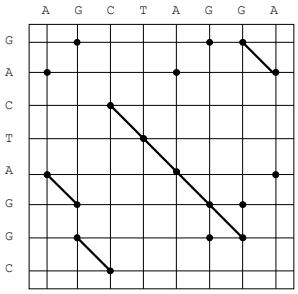
Visual Alignments (Dot Plots)

- Matrix
  - Rows: Characters in one sequence
  - Columns: Characters in second sequence
- Filling
  - Loop through each row; if character in row, col match, fill in the cell
  - Continue until all cells have been examined

54

The Dot Matrix

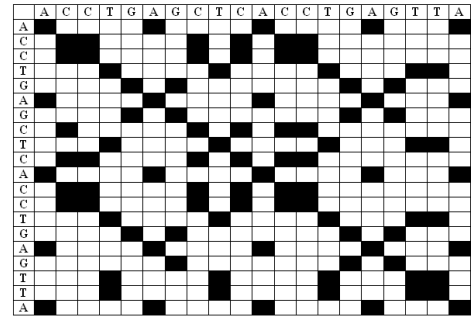
- established in 1970 by A.J. Gibbs and G.A.McIntyre
- method for comparing two amino acid or nucleotide sequences



- each sequence builds one axis of the grid
- one puts a dot, at the intersection of same letters appearing in both sequences
- scan the graph for a series of dots
  - reveals similarity
  - or a string of same characters
- longer sequences can also be compared on a single page, by using smaller dots

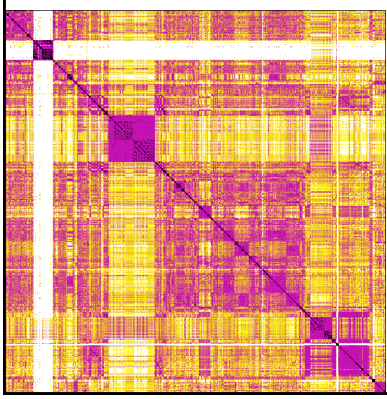
55

Example Dot Plot



56

An entire software module of a telecommunications switch; about two million lines of C

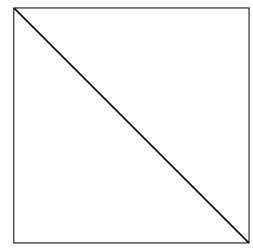


Darker areas indicate regions with a lot of matches (a high degree of similarity). Lighter areas indicate regions with few matches (a low degree of similarity). Dark areas along the main diagonal indicate sub-modules. Dark areas off the main diagonal indicate a degree of similarity between sub-modules The largest dark squares are formed by redundancies in initializations of signal-tables and finite-state machines.

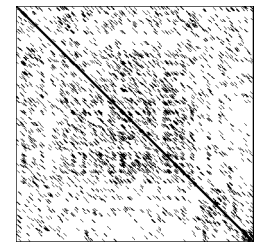
57

The Dot Matrix

The very stringent, self-dotplot:



The non-stringent self-dotplot:



58

Noise in Dot Plots

- Nucleic Acids (DNA, RNA)
  - 1 out of 4 bases matches at random
- Stringency (The condition under which a DNA sequence can bind to related or non-specific sequences. For example, high temperature and lower salt increases stringency such that non-specific binding or binding with low melting temperature will dissolve)
  - Window size is considered
  - Percentage of bases matching in the window is set as threshold

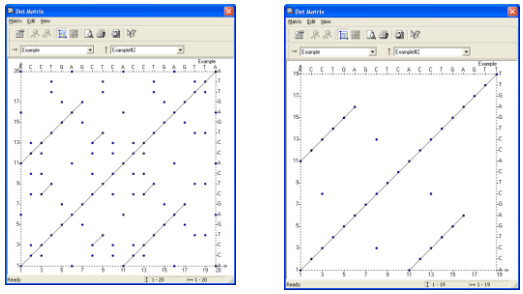
59

The Dot Matrix

- to filter out random matches, one uses sliding windows
- a dot is printed only if a minimal number of matches occur
- rule of thumb:
  - larger windows for DNAs (only 4 bases, more random matches)
  - typical window size is 15 and stringency of 10

60

Reduction of Dot Plot Noise

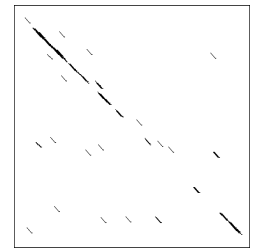


Self alignment of ACCTGAGCTCACCTGAGTTA

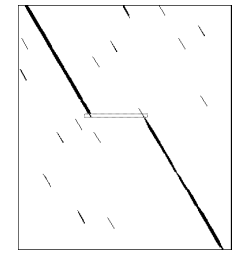
61

The Dot Matrix

Two similar, but not identical, sequences



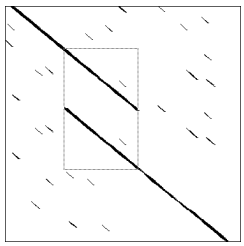
An indel (insertion or deletion):



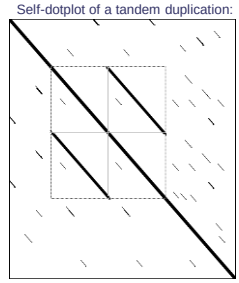
62

The Dot Matrix

A tandem duplication:



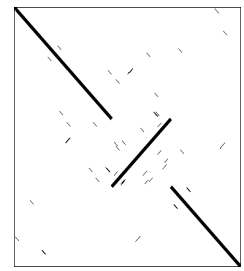
Self-dotplot of a tandem duplication:



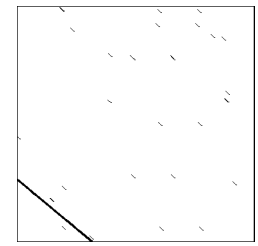
63

The Dot Matrix

An inversion:



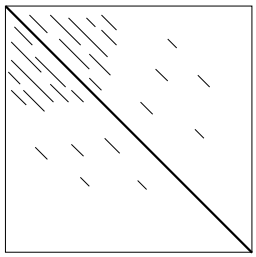
Joining sequences:



64

The Dot Matrix

Self dotplot with repeats:



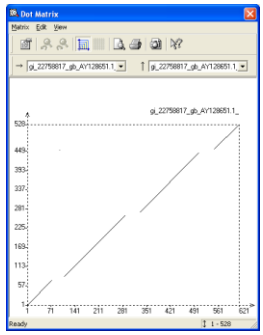
65

The Dot Matrix

- the dot matrix method reveals the presence of insertions or deletions
- comparing a single sequence to itself can reveal the presence of a repeat of a subsequence
  - Inverted repeats = reverse complement
    - Used to determine folding of RNA molecules
- self comparison can reveal several features:
  - similarity between chromosomes
  - tandem genes
  - repeated domains in a protein sequence
  - regions of low sequence complexity (same characters are often repeated)

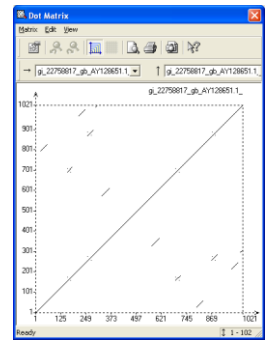
66

Insertions/Deletions



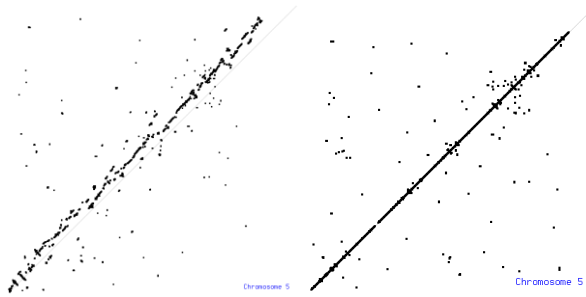
67

Repeats/Inverted Repeats



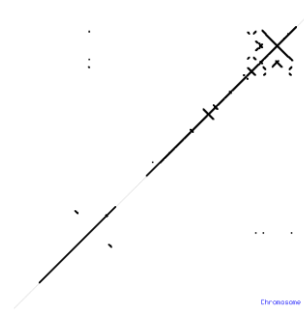
68

Comparing Genome Assemblies



69

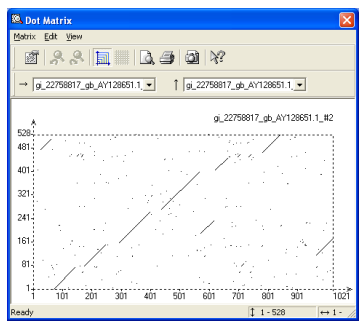
Chromosome Y self comparison



70

Available Dot Plot Programs

- Vector NTI software package (under AlignX)



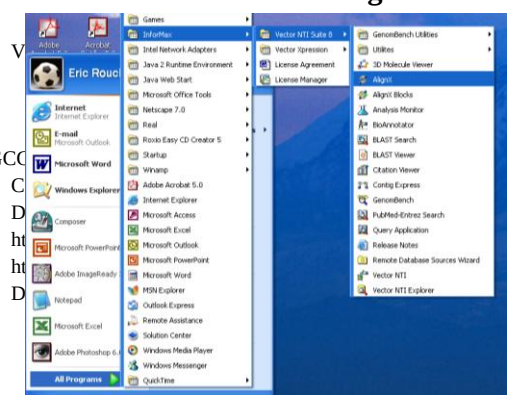
71

Available Dot Plot Programs

- V

GCC

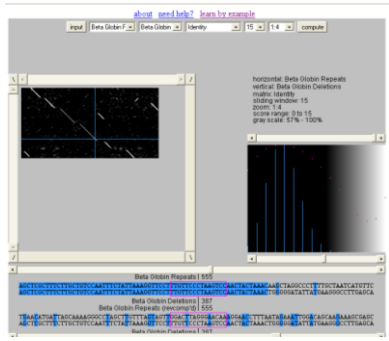
- C
- D
- ht
- ht
- D



72

Available Dot Plot Programs

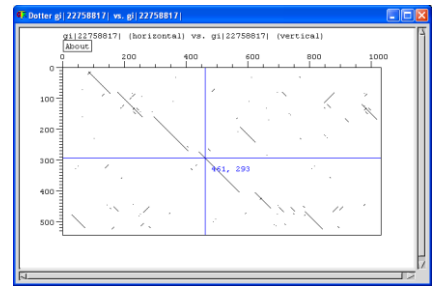
Dotlet (Java Applet)
http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html



73

Available Dot Plot Programs

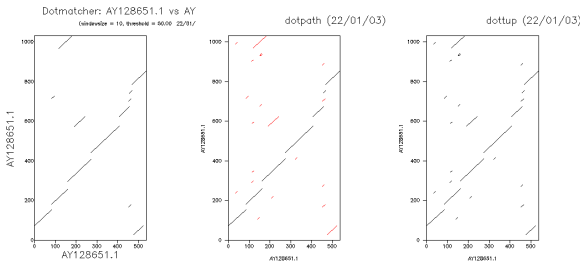
Dotter (http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html)



74

Available Dot Plot Programs

EMBOSS DotMatcher, DotPath,DotUp



75

Available Dot Plot Programs

GCG software package:

- Compare http://www.hku.hk/bruhk/gcgdoc/compare.html
- DotPlot+ http://www.hku.hk/bruhk/gcgdoc/dotplot.html

- DNA strider
- PipMaker

76

The Dot Matrix

- When to use the Dot Matrix method?
  - unless the sequences are known to be very much alike
- limits of the Dot Matrix
  - doesn't readily resolve similarity that is interrupted by insertion or deletions
  - Difficult to find the best possible alignment (optimal alignment)
  - most computer programs don't show an actual alignment

77

Dot Plot References

Gibbs, A. J. & McIntyre, G. A. (1970).
The diagram method for comparing sequences. its use with amino acid and nucleotide sequences.
Eur. J. Biochem. **16**, 1-11.

Staden, R. (1982).
An interactive graphics program for comparing and aligning nucleic-acid and amino-acid sequences.
Nucl. Acid. Res. **10** (9), 2951-2961.

78

Next step

We must define quantitative measures of sequence similarity and difference!

- *Hamming* distance:
  - *# of positions with mismatching characters*

```
AGTC
CGTA
```
Hamming distance = 2

- *Levenshtein* (or *edit*) distance:
  - *# of operations required to change one string into the other (deletion, insertion, substitution)*

```
AG-TCC
CGCTCA
```
Levenshtein distance = 3

79

Scoring

- +1 for a match -1 for a mismatch?
- should gaps be allowed?
  - if yes how should they be scored?
- what is the best algorithm for finding the optimal alignment of two sequences?
- is the produced alignment significant?

80

Determining Optimal Alignment

- Two sequences: X and Y
  - |X| = m; |Y| = n
  - Allowing gaps, |X| = |Y| = m+n
- Brute Force
- Dynamic Programming

81

Brute Force

- Determine all possible subsequences for X and Y
  - 2^{m+n} subsequences for X, 2^{m+n} for Y!
- Alignment comparisons
  - $2^{m+n} * 2^{m+n} = 2^{(2(m+n))} = 4^{m+n}$ comparisons
- Quickly becomes impractical

82

Dynamic Programming

- Used in Computer Science
- Solve optimization problems by dividing the problem into independent subproblems
- Sequence alignment has optimal substructure property
  - Subproblem: alignment of prefixes of two sequences
  - Each subproblem is computed once and stored in a matrix

83

Dynamic Programming

- Optimal score: built upon optimal alignment computed to that point
- Aligns two sequences beginning at ends, attempting to align all possible pairs of characters

84

Dynamic Programming

- Scoring scheme for matches, mismatches, gaps
- Highest set of scores defines optimal alignment between sequences
- Match score: DNA – exact match; Amino Acids – mutation probabilities
- Guaranteed to provide optimal alignment given:
  - Two sequences
  - Scoring scheme

85

Steps in Dynamic Programming

- Initialization
- Matrix Fill (scoring)
- Traceback (alignment)

DP Example:

Sequence #1: GAATTCAGTTA; M = 11

Sequence #2: GGATCGA; N = 7

- $s(a_i b_j)$ = +5 if $a_i = b_j$ (match score)
- $s(a_i b_j)$ = -3 if $a_i \neq b_j$ (mismatch score)
- w = -4 (gap penalty)

86

View of the DP Matrix

- M+1 rows, N+1 columns

	-	G	A	A	T	T	C	A	G	T	T	A
-												
G												
G												
A												
T												
C												
G												
A												

87

Global Alignment (Needleman-Wunsch)

- Attempts to align all residues of two sequences
- ***INITIALIZATION***: First row and first column set
- $S_{i,0} = w * i$
- $S_{0,j} = w * j$

88

Initialized Matrix(Needleman-Wunsch)

	-	G	A	A	T	T	C	A	G	T	T	A
-	0	-4	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44
G	-4											
G	-8											
A	-12											
T	-16											
C	-20											
G	-24											
A	-28											

89

Matrix Fill (Global Alignment)

$S_{i,j}$ = MAXIMUM[

$S_{i-1, j-1} + s(a_i, b_j)$ (match/mismatch in the diagonal),

$S_{i,j-1} + w$ (gap in sequence #1),

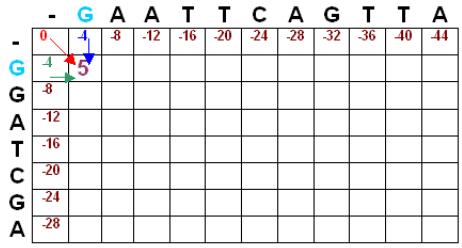
$S_{i-1,j} + w$ (gap in sequence #2)

]

90

Matrix Fill (Global Alignment)

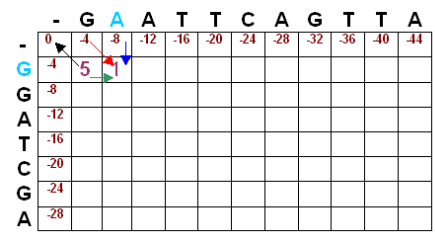
- $S_{1,1} = MAX[S_{0,0} + 5, S_{1,0} - 4, S_{0,1} - 4] = MAX[5, -8, -8]$



91

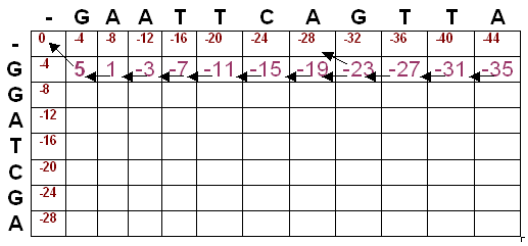
Matrix Fill (Global Alignment)

- $S_{1,2} = MAX[S_{0,1} -3, S_{1,1} - 4, S_{0,2} - 4] = MAX[-4 - 3, 5 – 4, -8 – 4] = MAX[-7, 1, -12] = 1$



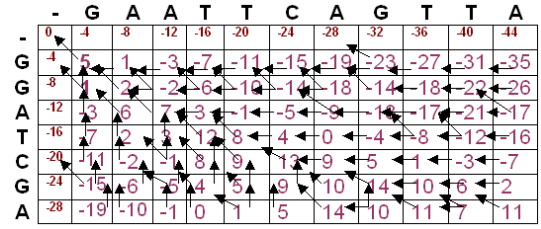
92

Matrix Fill (Global Alignment)



93

Filled Matrix (Global Alignment)



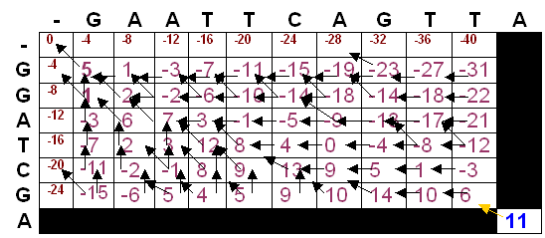
94

Trace Back (Global Alignment)

- maximum global alignment score = 11 (value in the lower right hand cell).
-
- Traceback begins in position $S_{M,N}$; i.e. the position where both sequences are globally aligned.
-
- At each cell, we look to see where we move next according to the pointers.

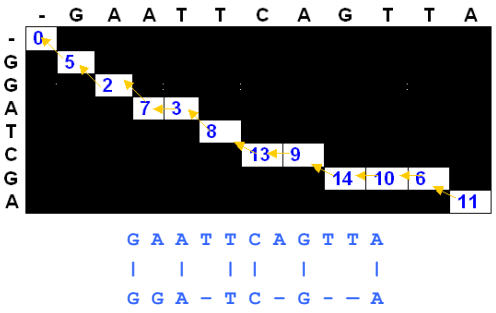
95

Trace Back (Global Alignment)



96

Global Trace Back




```
G A A T T C A G T T A
|   |   | |   |     |
G G A - T C - G - - A
```

97

Checking Alignment Score

```
G A A T T C A G T T A
|   |   | |   |     |
G G A - T C - G - - A

+ - + - + + - + - - +
5 3 5 4 5 5 4 5 4 4 5
```

5 – 3 + 5 – 4 + 5 + 5 – 4 + 5 – 4 – 4 + 5 = 11 ✓

98

Local Alignment

- Smith-Waterman: obtain highest scoring local match between two sequences
- Requires 2 modifications:
  - Negative scores for mismatches
  - When a value in the score matrix becomes negative, reset it to zero (begin of new alignment)

99

Local Alignment Initialization

- Values in row 0 and column 0 set to 0.

	-	G	A	A	T	T	C	A	G	T	T	A
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

100

Matrix Fill (Local Alignment)

$S_{i,j}$ = MAXIMUM [

$S_{i-1, j-1} + s(a_i, b_j)$ (match/mismatch in the diagonal),

$S_{i,j-1} + w$ (gap in sequence #1),

$S_{i-1,j} + w$ (gap in sequence #2),

0]

101

Matrix Fill (Local Alignment)

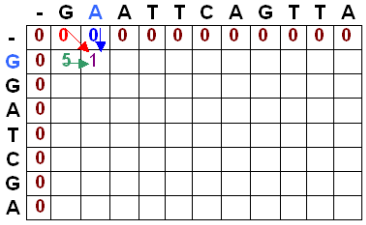
$S_{1,1} = MAX[S_{0,0} + 5, S_{1,0} - 4, S_{0,1} - 4, 0] = MAX[5, -4, -4, 0] = 5$

	-	G	A	A	T	T	C	A	G	T	T	A
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	5										
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

102

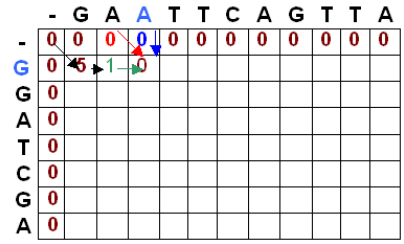
Matrix Fill (Local Alignment)

$S_{1,2} = MAX[S_{0,1} -3, S_{1,1} - 4, S_{0,2} – 4, 0] = MAX[0 - 3, 5 – 4, 0 – 4, 0] = MAX[-3, 1, -4, 0] = 1$

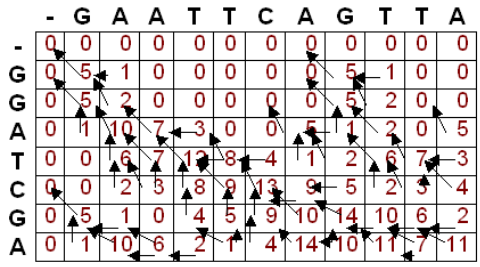


Matrix Fill (Local Alignment)

$S_{1,3} = MAX[S_{0,2} -3, S_{1,2} - 4, S_{0,3} – 4, 0] = MAX[0 - 3, 1 – 4, 0 – 4, 0] = MAX[-3, -3, -4, 0] = 0$



Filled Matrix (Local Alignment)



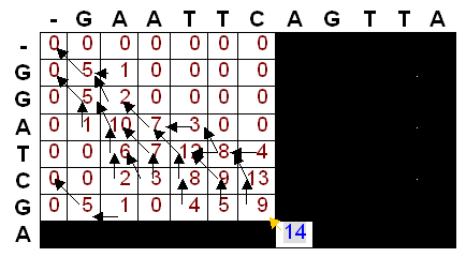
Trace Back (Local Alignment)

- maximum local alignment score for the two sequences is 14
- found by locating the highest values in the score matrix
- 14 is found in two separate cells, indicating multiple alignments producing the maximal alignment score

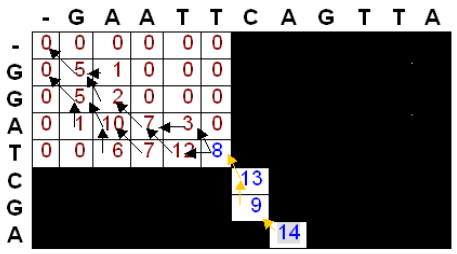
Trace Back (Local Alignment)

- Traceback begins in the position with the highest value.
- At each cell, we look to see where we move next according to the pointers
- When a cell is reached where there is not a pointer to a previous cell, we have reached the beginning of the alignment

Trace Back (Local Alignment)

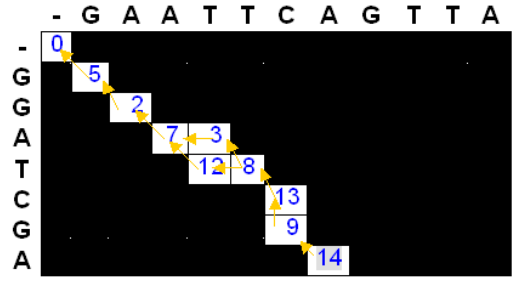


Trace Back (Local Alignment)



109

Trace Back (Local Alignment)



110

Maximum Local Alignment

```
G A A T T C - A      G A A T T C - A
|   | |   |  |      |   |   | |  |
G G A T - C G A      G G A - T C G A

+ - + + - + - +      + - + - + + - +
5 3 5 5 4 5 4 5      5 3 5 4 5 5 4 5
```

111

Overlap Alignment

Consider the following problem:

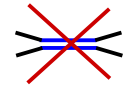
- Find the most significant overlap between two sequences?
- Possible overlap relations:

a.

b.

Difference from local alignment:

Here we require alignment between the endpoints of the two sequences.



Overlap Alignment

Initialization: $S_{i,0} = 0$, $S_{0,j} = 0$

Recurrence: as in global alignment

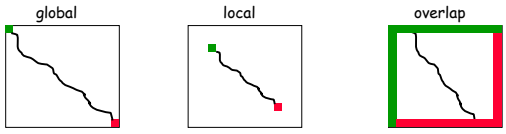
$S_{i,j}$ = **MAXIMUM [**

$S_{i-1, j-1} + s(a_i, b_j)$ (match/mismatch in the diagonal),

$S_{i,j-1} + w$ (gap in sequence #1),

$S_{i-1,j} + w$ (gap in sequence #2) **]**

Score: maximum value at the bottom line and rightmost line



Overlap Alignment - example

PAWHEAE
HEAGAWGHEE

Scoring scheme :
Match: +4
Mismatch: -1
Gap penalty: -5

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0										
A	0										
W	0										
H	0										
E	0										
A	0										
E	0										

Overlap Alignment

PAWHEAE
HEAGAWGHEE

Scoring scheme :
Match: +4
Mismatch: -1
Gap penalty: -5

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
A	0	-1									
W	0	-1									
H	0	4									
E	0	-1									
A	0	-1									
E	0	-1									

Overlap Alignment

PAWHEAE
HEAGAWGHEE

Scoring scheme:
Match: +4
Mismatch: -1
Gap penalty: -5

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
A	0	-1	-2	3	-2	3	-2	-2	-2	-2	-2
W	0	-1	-2	-2	2	-2	7	2	-3	-3	-3
H	0	4	-1	-3	-3	1	2	6	6	1	-4
E	0	-1	8	3	-2	-3	0	1	5	10	5
A	0	-1	3	12	7	2	-2	-1	0	5	9
E	0	-1	3	7	11	6	1	-3	-2	4	9

Overlap Alignment

The best overlap is:

P A W H E A E - - - - - -
- - - H E A G A W G H E E

Pay attention!
A different scoring scheme could yield a different result, such as:

Scoring scheme :
Match: +4
Mismatch: -1
Gap penalty: -2

- - - P A W- H E A E
H E A G A W G H E E -

Sequence Alignment Variants

- Global alignment (The Needelman-Wunsch Algorithm)
  - Initialization: $S_{i,0} = i*w$, $S_{0,j} = j*w$
  - Score: $S_{i,j} = \mathbf{MAX}\,[S_{i-1,\,j-1} + s(a_i, b_j),\ S_{i,j-1} + w,\ S_{i-1,j} + w]$
- Local alignment (The Smith-Waterman Algorithm)
  - Initialization: $S_{i,0} = 0$, $S_{0,j} = 0$
  - Score: $S_{i,j} = \mathbf{MAX}\,[S_{i-1,\,j-1} + s(a_i, b_j),\ S_{i,j-1} + w,\ S_{i-1,j} + w,\ \mathbf{0}]$
- Overlap alignment
  - Initialization: $S_{i,0} = 0$, $S_{0,j} = 0$
  - Score: $S_{i,j} = \mathbf{MAX}\,[S_{i-1,\,j-1} + s(a_i, b_j),\ S_{i,j-1} + w,\ S_{i-1,j} + w]$

Scoring Matrices

- match/mismatch score
  - Not bad for similar sequences
  - Does not show distantly related sequences
- Likelihood matrix
  - Scores residues dependent upon likelihood substitution is found in nature
  - More applicable for amino acid sequences

119

Parameters of Sequence Alignment

Scoring Systems:

- Each symbol pairing is assigned a numerical value, based on a symbol comparison table.

Gap Penalties:

- Opening: The cost to introduce a gap
- Extension: The cost to elongate a gap

120

DNA Scoring Systems -very simple

Sequence 1
Sequence 2

```
actaccagttcatttgatacttctcaaa
   | ||     |       ||
   taccattaccgtgttaactgaaaggacttaaagact
```

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

Match: 1
Mismatch: 0
Score = 5

121

Protein Scoring Systems

Sequence 1 PTHPLASKTQILPEDLASEDLTI
Sequence 2 PTHPLAGERAIGLARLAEEDFGM

T:G = -2
T:T = 5
Score = 48

Scoring matrix

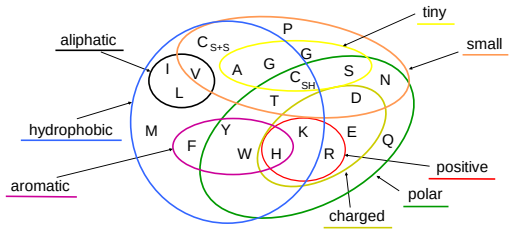
	C	S	T	P	A	G	N	D	.	.
C	9									
S	-1	4								
T	-1	1	5							
P	-3	-1	-1	7						
A	0	1	0	-1	4					
G	-3	0	-2	-2	0	6				
N	-3	1	0	-2	-2	0	5			
D	-3	0	-1	-1	-2	-1	1	6		
.										
.										

A scoring matrix is a table of values that describe the probability of a residue pair occurring in alignment.

122

Protein Scoring Systems

- Amino acids have different biochemical and physical properties that influence their relative replaceability in evolution.



123

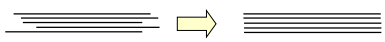
Protein Scoring Systems

- Scoring matrices reflect:
  - # of mutations to convert one to another
  - chemical similarity
  - observed mutation frequencies
- Log odds matrices:
  - the values are logarithms of probability ratios of the probability of an aligned pair to the probability of a random alignment.
- Widely used scoring matrices:
  - PAM
  - BLOSUM

124

PAM matrices (Percent Accepted Mutations)

- Derived from global alignments of protein families . Family members share at least 85% identity (Dayhoff *et al.*, 1978).



- Construction of phylogenetic tree and ancestral sequences of each protein family
- Computation of number of replacements for each pair of amino acids

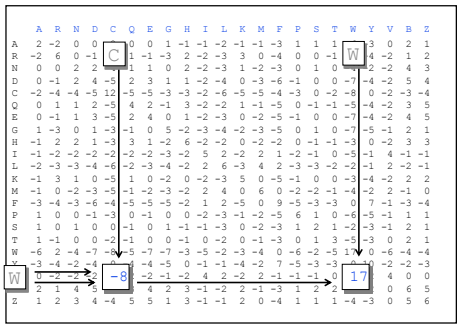
125

PAM matrices

- The numbers of replacements were used to compute a so-called PAM-1 matrix.
- The PAM-1 matrix reflects an average change of 1% of all amino acid positions.
- PAM matrices for larger evolutionary distances can be extrapolated from the PAM-1 matrix by multiplication.
- PAM250 = 250 mutations per 100 residues.
- Greater numbers mean bigger evolutionary distance

126

PAM 250



A value of 0 indicates the frequency of alignment is random

log(freq(observed)/freq(expected))

127

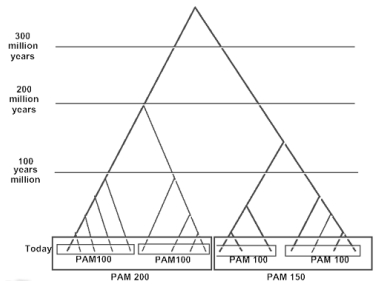
Amino Acid Frequency

freq(expected) = $f(AA_i)*f(AA_j)$

	1978	1991
L	0.085	0.091
A	0.087	0.077
G	0.089	0.074
S	0.070	0.069
V	0.065	0.066
E	0.050	0.062
T	0.058	0.059
K	0.081	0.059
I	0.037	0.053
D	0.047	0.052
R	0.041	0.051
P	0.051	0.051
N	0.040	0.043
Q	0.038	0.041
F	0.040	0.040
Y	0.030	0.032
M	0.015	0.024
H	0.034	0.023
C	0.033	0.020
W	0.010	0.014

128

Use Different PAM's for Different Evolutionary Distances

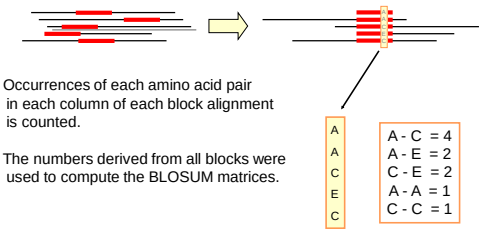


(Adapted from D Brutlag, Stanford)

129

BLOSUM (Blocks Substitution Matrix)

- Derived from alignments of domains of distantly related proteins (Henikoff & Henikoff,1992).



- Occurrences of each amino acid pair in each column of each block alignment is counted.
- The numbers derived from all blocks were used to compute the BLOSUM matrices.

A A C E C

A - C = 4
A - E = 2
C - E = 2
A - A = 1
C - C = 1

130

BLOSUM (Blocks Substitution Matrix)

- Sequences within blocks are clustered according to their level of identity.
- Clusters are counted as a single sequence.
- Different BLOSUM matrices differ in the percentage of sequence identity used in clustering.
- The number in the matrix name (e.g. 62 in BLOSUM62) refers to the percentage of sequence identity used to build the matrix.
- Greater numbers mean smaller evolutionary distance.

131

TIPS on choosing a scoring matrix

- Generally, BLOSUM matrices perform better than PAM matrices for local similarity searches (Henikoff & Henikoff, 1993).
- When comparing closely related proteins one should use lower PAM or higher BLOSUM matrices,
- For distantly related proteins higher PAM or lower BLOSUM matrices.
- For database searching the commonly used matrix is BLOSUM62.

132

Nucleic Acid Scoring Scheme

- Transition mutation (more common)
  - purine ⟷ purine A ⟷ G
  - pyrimidine ⟷ pyrimidine T ⟷ C
- Transversion mutation
  - purine ⟷ pyrimidine A, G ⟷ T, C

	A	G	T	C
A	20	10	5	5
G	10	20	5	5
T	5	5	20	10
C	5	5	10	20

133

Amino acid exchange matrices

Amino acids are **not** equal:

1. Some are easily substituted because they have similar:
   - physico-chemical properties
   - structure
2. Some mutations between amino acids occur more often due to similar codons

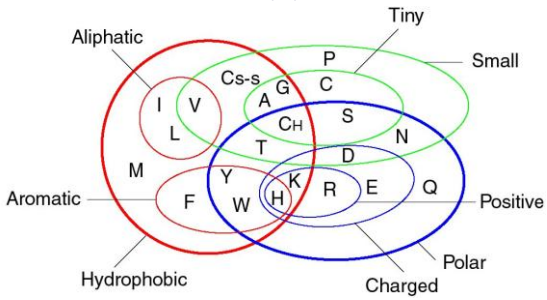
The two above observations give us ways to define *substitution matrices*

134

Properties of Amino Acids

Sequence similarity

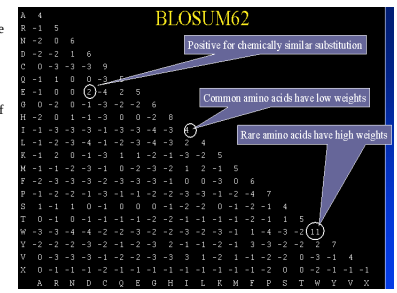
• substitutions with similar chemical properties



135

Scoring Matrices

- table of values that describe the probability of a residue pair occurring in an alignment
- the values are logarithms of ratios of two probabilities
  1. probability of random occurrence of an amino acid (diagonal)
  2. probability of meaningful occurrence of a pair of residues



136

Scoring Matrices

Widely used matrices

- **PAM** (Percent Accepted Mutation) / **MDM** (Mutation Data Matrix) / **Dayhoff**
  - Derived from *global* alignments of *closely* related sequences.
  - Matrices for greater evolutionary distances are extrapolated from those for lesser ones.
  - The number with the matrix (PAM40, PAM100) refers to the evolutionary distance; greater numbers are greater distances.
  - PAM-1 corresponds to about 1 million years of evolution
  - for distant (global) alignments, Blosum50, Gonnet, or (still) PAM250
- **BLOSUM** (Blocks Substitution Matrix)
  - Derived from *local, ungapped alignments of distantly related sequences*
  - *All matrices are directly calculated; no extrapolations are used*
  - *The number after the matrix (BLOSUM62) refers to the minimum percent identity of the blocks used to construct the matrix; greater numbers are lesser distances.*
  - *The BLOSUM series of matrices generally perform better than PAM matrices for local similarity searches.*
  - For local alignment, Blosum 62 is often superior
- ***Structure-based matrices***
- ***Specialized Matrices***

137

Scoring Matrices

The relationship between BLOSUM and PAM substitution matrices

- BLOSUM matrices with higher numbers and PAM matrices with low numbers are designed for comparisons of closely related sequences.
- BLOSUM matrices with low numbers and PAM matrices with high numbers are designed for comparisons of distantly related proteins.

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent ⟷		*More divergent*

http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Scoring2.html

138

Percent Accepted Mutation (PAM or Dayhoff) Matrices

- Studied by Margaret Dayhoff
- Amino acid substitutions
  - Alignment of common protein sequences
  - 1572 amino acid substitutions
  - 71 groups of protein, 85% similar
- "Accepted" mutations – do not negatively affect a protein's fitness
- Similar sequences organized into phylogenetic trees
- Number of amino acid changes counted
- Relative mutabilities evaluated
- 20 x 20 amino acid substitution matrix calculated

139

Percent Accepted Mutation (PAM or Dayhoff) Matrices

- PAM 1: 1 accepted mutation event per 100 amino acids; PAM 250: 250 mutation events per 100 …
- PAM 1 matrix can be multiplied by itself N times to give transition matrices for sequences that have undergone N mutations
- PAM 250: 20% similar; PAM 120: 40%; PAM 80: 50%; PAM 60: 60%

140

PAM1 matrix

normalized probabilities multiplied by 10000

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	1	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

141

Log Odds Matrices

- PAM matrices converted to log-odds matrix
  - Calculate odds ratio for each substitution
    - Taking scores in previous matrix
    - Divide by frequency of amino acid
  - Convert ratio to log10 and multiply by 10
  - Take average of log odds ratio for converting A to B and converting B to A
  - Result: Symmetric matrix
  - EXAMPLE: Mount pp. 80-81

142

PAM250 Log odds matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-5	-4	-5	-4	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	-3	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

143

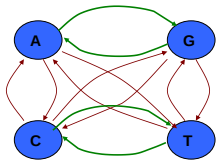
Blocks Amino Acid Substitution Matrices (BLOSUM)

- Larger set of sequences considered
- Sequences organized into signature blocks
- Consensus sequence formed
  - 60% identical: BLOSUM 60
  - 80% identical: BLOSUM 80

144

DNA Mutations

In addition to using a match/mismatch scoring scheme for DNA sequences, nucleotide mutation matrices can be constructed as well. These matrices are based upon two different models of nucleotide evolution: the first, the Jukes-Cantor model, assumes there are uniform mutation rates among nucleotides, while the second, the Kimura model, assumes that there are two separate mutation rates: one for transitions (where the structure of purine/pyrimidine stays the same), and one for transversions. Generally, the rate of transitions is thought to be higher than the rate of transversions.



PURINES: A, G
PYRIMIDINES C, T

Transitions: A↔G; C↔T

Transversions: A↔C, A↔T, C↔G, G↔T

Nucleic Acid Scoring Matrices

- Two mutation models:
  - Jukes-Cantor Model of evolution: α = common rate of base substitution
  - Kimura Model of Evolution: α = rate of transitions; β = rate of transversions
    - Transitions
    - Transversions

$$R = \begin{array}{c} \\ A \\ C \\ G \\ U \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & U \end{array} \\ \begin{pmatrix} * & 0.25\alpha & 0.25\alpha & 0.25\alpha \\ 0.25\alpha & * & 0.25\alpha & 0.25\alpha \\ 0.25\alpha & 0.25\alpha & * & 0.25\alpha \\ 0.25\alpha & 0.25\alpha & 0.25\alpha & * \end{pmatrix} \end{array}$$

$$R = \begin{array}{c} \\ A \\ C \\ G \\ U \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & U \end{array} \\ \begin{pmatrix} * & 0.25\beta & 0.25\alpha & 0.25\beta \\ 0.25\beta & * & 0.25\beta & 0.25\alpha \\ 0.25\alpha & 0.25\beta & * & 0.25\beta \\ 0.25\beta & 0.25\alpha & 0.25\beta & * \end{pmatrix} \end{array}$$

Nucleotide substitution matrices with the equivalent distance of 1 PAM

A. Model of uniform mutation rates among nucleotides.

	A	G	T	C
A	0.99			
G	0.00333	0.99		
T	0.00333	0.00333	0.99	
C	0.00333	0.00333	0.00333	0.99

B. Model of 3-fold higher transitions than transversions.

	A	G	T	C
A	0.99			
G	0.006	0.99		
T	0.002	0.002	0.99	
C	0.002	0.002	0.006	0.99

Nucleotide substitution matrices with the equivalent distance of 1 PAM

A. Model of uniform mutation rates among nucleotides.

	A	G	T	C
A	2			
G	-6	2		
T	-6	-6	2	
C	-6	-6	-6	2

B. Model of 3-fold higher transitions than transversions.

	A	G	T	C
A	2			
G	-5	2		
T	-7	-7	2	
C	-7	-7	-5	2

Linear vs. Affine Gaps

- The scoring matrices used to this point assume a linear gap penalty where each gap is given the same penalty score.
- However, over evolutionary time, it is more likely that a contiguous block of residues has become inserted/deleted in a certain region (for example, it is more likely to have 1 gap of length k than k gaps of length 1).
- Therefore, a better scoring scheme to use is an initial higher penalty for opening a gap, and a smaller penalty for extending the gap.

Linear vs. Affine Gaps

- Gaps have been modeled as linear
- More likely contiguous block of residues inserted or deleted
  - 1 gap of length k rather than k gaps of length 1
- Scoring scheme should penalize new gaps more

Affine Gap Penalty

- $w_x = g + r(x-1)$

- w_x : total gap penalty; g: gap open penalty; r: gap extend penalty; x: gap length
-
- gap penalty chosen relative to score matrix
  - Gaps not excluded
  - Gaps not over included
  - Typical Values: g = -12; r = -4

151

Affine Gap Penalty and Dynamic Programming

$$M_{i,j} = \max \{ D_{i-1,j-1} + \text{subst}(A_i, B_j) , M_{i-1,j-1} + \text{subst}(A_i, B_j) , I_{i-1,j-1} + \text{subst}(A_i, B_j) \}$$

$$D_{i,j} = \max \{ D_{i,j-1} - \text{extend}, \quad M_{i,j-1} - \text{open} \}$$

$$I_{i,j} = \max \{ M_{i-1,j} - \text{open}, \quad I_{i-1,j} - \text{extend} \}$$

where M is the match matrix, D is delete matrix, and I is insert matrix

152

Drawbacks to DP Approaches

- Dynamic programming approaches are guaranteed to give the optimal alignment between two sequences given a scoring scheme.
- However, the two main drawbacks to DP approaches is that they are compute and memory intensive, in the cases discussed to this point taking at least $O(n^2)$ space, between $O(n^2)$ and $O(n^3)$ time.
- Linear space algorithms have been used in order to deal with one drawback to dynamic programming. The basic idea is to concentrate only on those areas of the matrix more likely to contain the maximum alignment. The most well-known of these linear space algorithms is the Myers-Miller algorithm. Compute intensive

153

Alternative DP approaches

- Linear space algorithms Myers-Miller
- Bounded Dynamic Programming

- Ewan Birney's Dynamite Package
  - Automatic generation of DP code

154

Assessing Significance of Alignment

- When two sequences of length m and n are not obviously similar but show an alignment, it becomes necessary to assess the significance of the alignment. The alignment of scores of random sequences has been shown to follow a Gumbel extreme value distribution.

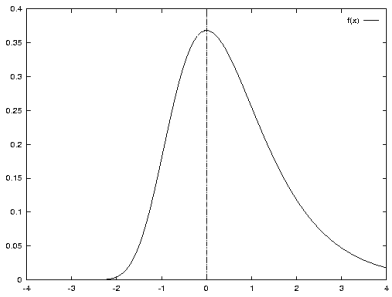
155

Significance of Alignment

- Determine probability of alignment occurring at random
  - Sequence 1: length m
  - Sequence 2: length n
- Random sequences:
  - When two sequences of length m and n are not obviously similar but show an alignment, it becomes necessary to assess the significance of the alignment.
  - The alignment of scores of random sequences has been shown to follow a Gumbel extreme value distribution.

156

Gumbel Extreme Value Distribution



- http://roso.epfl.ch/mbi/papers/discretechoice/node11.html
- http://mathworld.wolfram.com/GumbelDistribution.html
- http://en.wikipedia.org/wiki/Generalized_extreme_value_distribution

157

Probability of Alignment Score

- Using a Gumbel extreme value distribution, the expected number of alignments with a score at least S (E-value) is:

$$E = Kmn\, e^{-\lambda S}$$

  - m,n: Lengths of sequences
  - *K* ,λ: statistical parameters dependent upon scoring system and background residue frequencies

158

- Recall that the log-odds scoring schemes examined to this point normally use a $S = 10*\log_{10}x$ scoring system.
- We can normalize the raw scores obtained using these non-gapped scoring systems to obtain the amount of bits of information contained in a score (or the amount of nats of information contained within a score).

159

Converting to Bit Scores

A raw score can be normalized to a bit score using the formula:

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

- The E-value corresponding to a given bit score can then be calculated as:

$$E = mn\, 2^{-S'}$$

160

- Converting to nats is similar. However, we just substitute e for 2 in the above equations. Converting scores to either bits or nats gives a standardized unit by which the scores can be compared.

161

P-Value

- P values can be calculated as the probability of obtaining a given score at random. P-values can be estimated as:

$$P = 1 - e^{-E}$$

which is approximately e^{-E}

162

A quick determination of significance

- If a scoring matrix has been scaled to bit scores, then it can quickly be determined whether or not an alignment is significant.
- For a typical amino acid scoring matrix, K = 0.1 and lambda depends on the values of the scoring matrix.
- If a PAM or BLOSUM matrix is used, then lambda is precomputed.
- For instance, if the log odds matrix is in units of bits, then lambda = $\log_e 2$, and the significance cutoff can be calculates as $\log_2(mn)$.

163

Significance of Ungapped Alignments

- PAM matrices are $10 * \log_{10} x$
- Converting to $\log_2 x$ gives bits of information
- Converting to $\log_e x$ gives nats of information

Quick Calculation:

- If bit scoring system is used, significance cutoff is:

$$\log_2(mn)$$

164

Example

- Suppose we have two sequences, each approximately 250 amino acids long that are aligned using a Smith-Waterman approach.
- Significance cutoff is:
  - $\log_2(250 * 250) = 16$ bits

165

Example

- Using PAM250, the following alignment is found:

- F W L E V E G N S M T A P T G
- F W L D V Q G D S M T A P A G

166

Example

- Using PAM250, the score is calculated:

- F W L E V E G N S M T A P T G
- F W L D V Q G D S M T A P A G

- S = 9 + 17 + 6 + 3 + 4 + 2 + 5 + 2 + 2 + 6 + 3 + 2 + 6 + 1 + 5 = 73

167

PAM250 matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	−2	1	3																		T
P	−3	1	0	6																	P
A	−2	1	1	1	2																A
G	−3	1	0	−1	1	5															G
N	−4	1	0	−1	0	0	2														N
D	−5	0	0	−1	0	1	2	4													D
E	−5	0	0	−1	0	0	1	3	4												E
Q	−5	−1	−1	0	0	−1	1	2	2	4											Q
H	−3	−1	−1	0	−1	−2	2	1	1	3	6										H
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6									R
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5								K
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6							M
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5						I
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6					L
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4				V
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9			F
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10		Y
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17	W
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

168

Significance Example

- S is in 10 * $\log_{10}$x, so this should be converted to a bit score:

- S = 10 $\log_{10}$x
- S/10 = $\log_{10}$x
- S/10 = $\log_{10}$x * ($\log_2$10/$\log_2$10)
- S/10 * $\log_2$10 = $\log_{10}$x / $\log_2$10
- S/10 * $\log_2$10 = $\log_2$x
- 1/3 S ~ $\log_2$x

- S' ~ 1/3S

169

Significance Example

- S' = 1/3S = 1/3 * 73 = 24.333 bits

- The significance cutoff is:
  log2(mn) = log2(250 * 250) = 16 bits

- Since the alignment score is above the significance cutoff, this is a significant local alignment.

170

Estimation of E and P

- When a PAM250 scoring matrix is being used, K is estimated to be 0.09, while lambda is estimated to be 0.229.
- For PAM250, K = 0.09; λ = 0.229

- We can convert the score to a bit score as follows:
- $S' = \lambda S - \ln Kmn$
- S' = 0.229 * 73 – ln 0.09 * 250 * 250
- S' = 16.72 – 8.63 = 8.09 bits
- $P(S' >= x) = 1 - e^{(-e^{-x})}$
- $P(S' >= 8.09) = 1 - e^{(-e^{-8.09})} = 3.1 * 10^{-4}$

- Therefore, we see that the probability of observing an alignment with a bit score greater than 8.09 is about 3 in 1000.

171

Significance of Gapped Alignments

- Gapped alignments make use of the same statistics as ungapped alignments in determining the statistical significance.
- However, in gapped alignments, the values for λ and K cannot be easily estimated.
- Empirical estimations and gap scores have been determined by looking at the alignments of randomized sequences.

172

Bayesian Statistics

- Bayesian statistics are built upon conditional probabilities,
  - which are used to derive the joint probability of two events or conditions.
- P(B|A) is the probability of B given condition A is true.
- P(B) is the probability of condition B occurring, regardless of conditions A.
- P(A, B): Joint probability of A and B occurring simultaneously

173

Bayesian Statistics

- Suppose that A can have two states, A1 and A2, and B can have two states, B1 and B2.
- Suppose that P(B1) = 0.3 is known.
- Therefore, P(B2) = 1 – 0.3 = 0.7.
- These probabilities are known as ***marginal probabilities***.
- Now we would like to determine the probability of A1 and B1 occurring together, which is denoted as: P(A1, B1) and is called the joint probability

174

Joint Probabilities

- Note that in this case the marginal probabilities A1 and A2 are missing. Thus, there is not enough information at this point to calculate the marginal probability.
- However, if more information about the joint occurrence of A1 and B1 are given, then the joint probabilities may be derived using Bayes Rule:
  - P(A1,B1) = P(B1)P(A1|B1)
  - P(A1,B1) = P(A1)P(B1|A1)

175

Bayesian Example

- Suppose that we are given P(A1|B1) = 0.8.
- Then, since there are only two different possible states for A,
- P(A2|B1) = 1 – 0.8 = 0.2.
- If we are also given P(A2|B2) = 0.7,
- then P(A1|B2) = 0.3.
- Using Bayes Rule, the joint probability of having states A1 and B1 occurring at the same time is
- P(B1)P(A1|B1) = 0.3 * 0.8 = 0.24 and
- P(A2,B2) = P(B2)P(A2|B2) = 0.7 * 0.7 = 0.49.
- The other joint probabilities can be calculated from these as well.

176

Posterior Probabilities

- Calculation of joint probabilities results in posterior probabilities
  - Not known initially
  - Calculated using
    - Prior probabilities
    - Initial information

177

Applications of Bayesian Statistics

- Evolutionary distance between two sequences
- Sequence Alignment
- Significance of Alignments
- Gibbs Sampling

178

Pairwise Sequence Alignment Programs

- needle
  - Global Needleman/Wunsch alignment
- water
  - Local Smith/Waterman alignment
- Blast 2 Sequences
  - NCBI
  - word based sequence alignment
- LALIGN
  - FASTA package
  - Mult. Local alignments

179

Various Sequence Alignments

Wise2 -- Genomic to protein

Sim4 -- Aligns expressed DNA to genomic sequence

spidey -- aligns mRNAs to genomic sequence

est2genome -- aligns ESTs to genomic sequence

180