

# Bioinformatics and sequence analysis

**Michael Nilges**  
Unité de Bio-Informatique Structurale  
Institut Pasteur, Paris  
Mars 2002

## Overview

2

- **Bioinformatics: a brief overview**
- **Organising knowledge: databanks and databases**
- **Protein sequence analysis**
  - Sequence alignment
  - Multiple alignment and sequence profiles
  - Phylogenetic trees

## I. Bioinformatics - a brief overview

## What is it?

4

### **Bioinformatics :**

Deduction of knowledge by computer analysis  
of biological data.

or: see 20000 pages on this issue on the WWW

## The data

5

- information stored in the genetic code (DNA)
- protein sequences
- 3D structures
- experimental results from various sources
- patient statistics
- scientific literature

## Algorithmic developments

6

- Important part of research in bioinformatics:  
methods for
  - data storage
  - data retrieval
  - data analysis

## Interdisciplinary research

- rapidly developing branch of biology
- highly interdisciplinary:  
using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics.
- many practical applications in biology and medicine.

Bioinformatics lecture March 5, 2002



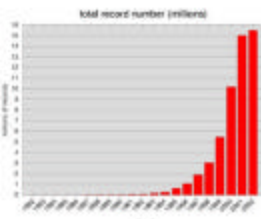
## Computation in biology...

- similar to other sciences:  
computational physics, computational chemistry  
derivation of physics laws from astronomical data
- already in the '20s biologists wanted to derive knowledge by induction
- reasons for recent development:  
development of computers and networks  
availability of data (sequences, 3D structures)  
amount of data

Bioinformatics lecture March 5, 2002



## Why?



- An avalanche of data:
  - Sequences
  - Function related
  - Structures
- requires computational approaches

Bioinformatics lecture March 5, 2002



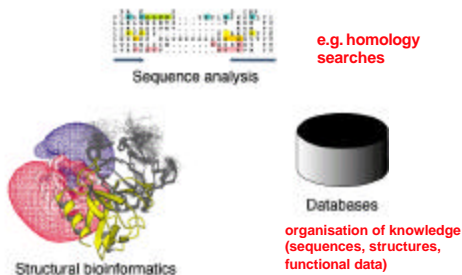
## Genomics

- New way to perform experiments:
  - accumulation of data:
    - sequences
    - structures,
    - function-related
  - not hypothesis-driven
- Hypothesis formed later and tested *in silico*

Bioinformatics lecture March 5, 2002



## Bioinformatics key areas



Bioinformatics lecture March 5, 2002



## Structural Bioinformatics

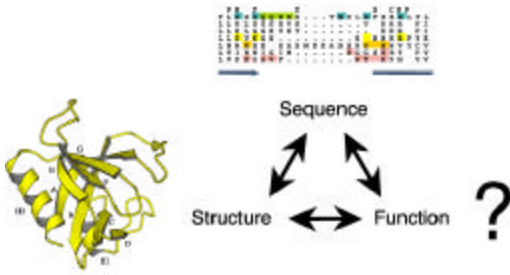
- Prediction of structure from sequence
  - secondary structure
  - homology modelling, threading
  - ab initio 3D prediction
- Analysis of 3D structure
  - structure comparison/ alignment
  - prediction of function from structure
  - molecular mechanics/ molecular dynamics
  - prediction of molecular interactions, docking
- Structure databases (RCSB)

Bioinformatics lecture March 5, 2002



## Structural Bioinformatics

13



Bioinformatics lecture March 5, 2002

EMBL

## II. Databases

### Organizing knowledge in databanks and databases

15

- Introduction
- Sequence databanks and databases
  - EMBL, SwissProt, TREMBL
- SRS: Sequence Retrieval system
- 3D structure database: the RCSB - PDB
- Domain databases

Bioinformatics lecture March 5, 2002

EMBL

### Biological databanks and databases

16

- Very fast growth of biological data
- Diversity of biological data:
  - primary sequences
  - 3D structures
  - functional data
- Database entry usually required for publication
  - Sequences
  - Structures
- Database entry may replace primary publication
  - genomic approaches

Bioinformatics lecture March 5, 2002

EMBL

### DNA sequence data bases

17

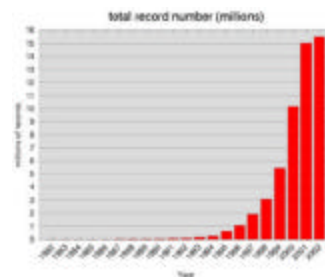
- Three databanks exchange data on a daily basis
- Data can be submitted and accessed at either location
- Genebank
  - [www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html](http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html)
- EMBL
  - [www.ebi.ac.uk/embl/index.html](http://www.ebi.ac.uk/embl/index.html)
- DNA DataBank of Japan (DDBJ)
  - [www.nig.ac.jp/home.html](http://www.nig.ac.jp/home.html)

Bioinformatics lecture March 5, 2002

EMBL

### EMBL database growth

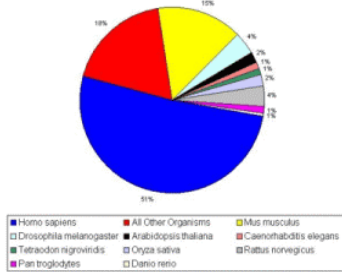
18



Bioinformatics lecture March 5, 2002

EMBL

## Distribution of entries



Bioinformatics lecture March 5, 2002



## EMBL database documentation

- Information on user manual
- release notes
- feature table definition... see
- <http://www.ebi.ac.uk/embl/Documentation>

Bioinformatics lecture March 5, 2002



## EMBL entry for insulin receptor

```

ID HSINR24 standard; DNA; PRI; 873 BP.
AC M32972;
DT 10-31-1990 [Rel. 24, Created]
DY 05-08-1992 [Rel. 33, Last updated, Version 4]
DE Human insulin receptor (hINSR) gene, exon 21.
KW insulin receptor.
OG Homo sapiens (human).
OC Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
OC Theria; Eutheria; Primates; Haplorhini; Catarrhini; Hominoidea.
RN [1]
RP 1-873
RA Seino S., Seino W., Bell G.I.;
RT "Human insulin-receptor gene";
RL Diabetes 19:123-128(1990).
CC Draft entry and computer-readable sequence for [1] kindly submitted
CC by G.I.Bell, 14-08-1990.
FH Key Location/Qualifiers
  
```

Bioinformatics lecture March 5, 2002



## EMBL entry 2: features

```

FT CDS             join(M32100:1824..1975,M32820:174..725,M32874:314..435,
FT                M32820:318..446,M32826:108..332,M32827:189..403,M32828:277
FT                ..402,M32829:126..374,M32830:106..273,M32831:187..388,
FT                M32832:122..158,M32833:161..435,M32834:92..223,M32835:95..
FT                244,M32836:92..194,M32837:281..138,M32838:174..380,M32839:
FT                117..227,M32840:85..204,M32861:124..244,M32842:101..220,60
FT                ..427)
FT /note="Human insulin receptor precursor"
FT source          1..873
FT /organism="Homo sapiens"
FT mol_peptide     join(M32100:1824..1975,M32820:174..725,M32874:314..420,
FT                M32820:318..446,M32830:106..332,M32831:189..403,M32832:277
FT                ..402,M32829:126..374,M32830:106..273,M32831:187..388,
FT                M32832:122..158,M32833:161..435,M32834:92..222,M32835:95..
FT                244,M32836:92..194,M32837:182..218,M32838:174..380,M32839:
FT                117..227,M32840:85..204,M32861:124..244,M32842:101..220,60
FT                ..424)
FT /note="Human insulin receptor"
FT gene_location  1..873
FT /note="cDNA: mRNA and introns"
FT locus          "3..91
FT /note="cDNA: Intron 1"
  
```

Bioinformatics lecture March 5, 2002



## EMBL entry 3: sequence

```

8Q  Sequence 873 BP; 399 A; 217 C; 234 G; 223 T; 0 other; 123
  GAGTGGTCC TCTGGGCGG AGGAGTGGC CAGTGGGCA GTAGTGGCA TCCAGCCCA 123
  AGTATGAGC AACCTCTCCG GAGTGTGCA ACTGTGCA GGCAGCTCG AACCTAGCT 183
  TCCAGTGGT GTGTGTCTC CAGCTGGGG AGTCAAGCG TCCAGTGGT GAGTGTGCG 243
  ...
  CCCCAGCCG CCCCAGCAG ATGAGAGCA AGCACTGTT TCCAGAGT CTTTCTTT 283
  TTTTCTTT TTTTCTTT CTGTTCTG AGTCTAGT TAAAGACA AACCTCTGT 343
  TGTGGGCC AATTTCGCA AGGAGAGCC 888
  //
  
```

Bioinformatics lecture March 5, 2002



## SwissProt protein sequence data base TREMBL translated EMBL

- hosted jointly by EBI (European Bioinformatics Institute, an EMBL outpost in Hinxton, UK) and SIB (Swiss Institute for Bioinformatics in Lausanne and Geneva)
- SwissProt is curated (Amos Bairoch)
  - quality checks
  - annotations
  - links to other databases
- TREMBL: automatic translation of EMBL automatic annotations

Bioinformatics lecture March 5, 2002









## Results of InterPro search for spectrin

InterPro search results for spectrin. The table shows various domain hits, with 'Spectrin repeat' highlighted by a red arrow.

Name	E-value	Description
Protein kinase domain	1.0e-10	Protein kinase domain
...	...	...
<b>Spectrin repeat</b>	1.0e-10	Spectrin repeat
...	...	...

## Spectrin repeat

Detailed view of the Spectrin repeat domain. The page includes a 3D ribbon diagram of the domain structure and a list of references.

## SMART database

SMART database search results for spectrin. The table shows various domain hits, with 'Spectrin repeat' highlighted by a red arrow.

Name	E-value	Description
...	...	...
<b>Spectrin repeat</b>	1.0e-10	Spectrin repeat
...	...	...

## Domain architecture of spectrin beta chain

Domain architecture of the spectrin beta chain. The diagram shows a linear map of the protein with colored boxes representing different domains, including the Spectrin repeat.

## Pfam home page

Pfam home page showing search options and domain family information. A 3D ribbon diagram of a protein structure is also visible.

## Compilations of links to databases

- at Institut Pasteur
  - [www.pasteur.fr/recherche/banques](http://www.pasteur.fr/recherche/banques)
- at Infobiogen (Evry)
  - [www.infobiogen.fr/services/deambulun/fr](http://www.infobiogen.fr/services/deambulun/fr)
- European bioinformatics institute (ebi)
  - [www.ebi.ac.uk/Databases/index.html](http://www.ebi.ac.uk/Databases/index.html)
- at the swiss institute for bioinformatics (SIB)
  - [www.expasy.org](http://www.expasy.org)
  - [www.expasy.org/alinks.html#Proteins](http://www.expasy.org/alinks.html#Proteins)

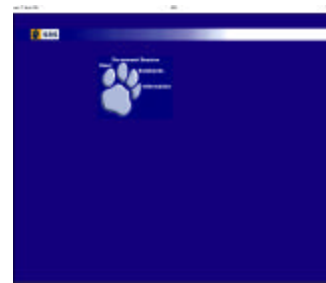
## SRS sequence retrieval system

- unified way to access and *link* information in different databases
- powerful queries
- launch applications (e.g. blast, clustalw...)
- temporary and permanent projects
- can be reached from the pasteur databank page
- [srs.pasteur.fr/cgi-bin/srs6/wgetz](http://srs.pasteur.fr/cgi-bin/srs6/wgetz)

Bioinformatics lecture March 5, 2002

INSTITUT PASTEUR  
UNIVERSITÉ PARIS 6

## SRS 6 start page



Bioinformatics lecture March 5, 2002

INSTITUT PASTEUR  
UNIVERSITÉ PARIS 6

## SRS access to databases

Bioinformatics lecture March 5, 2002

INSTITUT PASTEUR  
UNIVERSITÉ PARIS 6

## SRS quick search

Bioinformatics lecture March 5, 2002

INSTITUT PASTEUR  
UNIVERSITÉ PARIS 6

## SRS queries

- queries by simple words
- extension of words by wildcards
- linked by logical operators (and, or, &, ...)
- standard query form has 4 entry fields
- display list can be customized

Bioinformatics lecture March 5, 2002

INSTITUT PASTEUR  
UNIVERSITÉ PARIS 6

## Standard SRS query

Bioinformatics lecture March 5, 2002

INSTITUT PASTEUR  
UNIVERSITÉ PARIS 6



## GCG sequence format

### GCG format

```
PROTEIN of: xfstec check: 8509 from 1 to: 2256
<---No Config Comments--->
xfstec.seq Length: 2256 May 28, 1994 17:10 Type: N Check: 8509 ..
  1  SCTAGAAATG AAGAAATAAT CTTCTAATC CTTGAACACA TCTCTGGAGA
 51  TTTATCTCAG TTCTCAAGT GTTCAATTT CCGGACAGC GGTTCGCAA
101  AGTTCGAAA CCGTAGAG AGTGAATC AGAATTCAG CTTATCTCC
151  ATGATATTT CCGACGATG TACCGAAG AAGGATTTG CTTGGCGCA
201  ACCAGGTTG ATATCAACA ACCTATCTT GTTATTTAG TTTCGAAAA
251  CCGTAGCAA CCGCTATGT TAATCAATC AGAGCTTTA GAAAAGCG
```

Bioinformatics lecture March 5, 2002

UNIVERSITY OF  
SHEFFIELD

## GCG database format

### GCG database format

```
LOCUS   HME110   1601 bp  ss-sRNA   PR2   15-FEB-1993
... Séquence GDS de Genbank...
ORIGIN
857427 Length: 1601 January 7, 1994 17:29 Type: N Check: 9732 ..
  1  AAGAGACAG ACAGACTTGT AAGAGAGCC ATGAGAGCT
 51  CTTATGCTG GTTCTCTTNA CTTGGATTAAG ACCAGACCA G
```

- comments: up to "..."
- signal line with identifier "Check ...."
- sequence

Bioinformatics lecture March 5, 2002

UNIVERSITY OF  
SHEFFIELD

## Format conversions

- in GCG: specific command to convert from different formats (e.g., fromstaden)
- readseq:
- general conversion program
- available on [www at pasteur](http://www.pasteur.fr)

Bioinformatics lecture March 5, 2002

UNIVERSITY OF  
SHEFFIELD

## Protein sequence alignment (DNA alignment is analogous)

- Local sequence comparison:
- assumption of evolution by point mutations
  - amino acid replacement (by base replacement)
  - amino acid insertion
  - amino acid deletion
- scores:
  - positive for identical or similar
  - negative for different
  - negative for insertion in one of the two sequences

Bioinformatics lecture March 5, 2002

UNIVERSITY OF  
SHEFFIELD

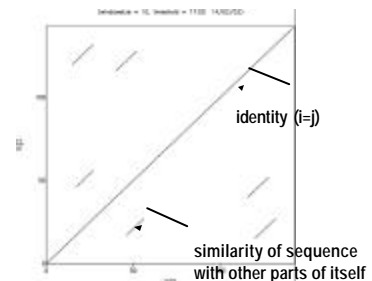
## Comparing two sequences: DotPlot

- Simple comparison without alignment
- Similarities between sequences show up in 2D diagram

Bioinformatics lecture March 5, 2002

UNIVERSITY OF  
SHEFFIELD

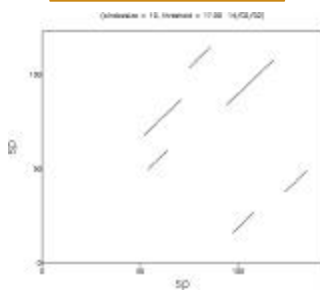
## Dotplot for a small protein against itself



Bioinformatics lecture March 5, 2002

UNIVERSITY OF  
SHEFFIELD

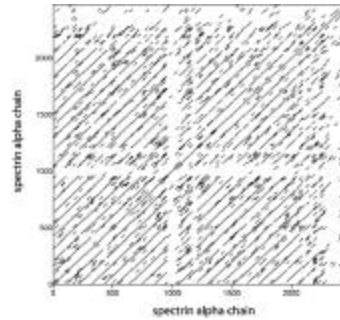
## Dotplot for two remotely homologous proteins



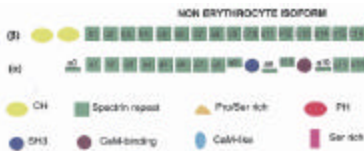
Bioinformatics lecture March 5, 2002



## Dotplot for protein with internal repeats



## Spectrin domain structure



Bioinformatics lecture March 5, 2002



## 3 alignments of globin sequences: right or wrong?

```
HBA_HUMAN GSAQVKGHGKVKVADALTNVAHVVDDMPNALSALSOLHAEKL
G+ +VK+HGKKV A++++AH+D++ +++++LS+LH KL
HBB_HUMAN GNPVKVARGKKVLPAPSGLAHLONLKGTFATLSLHAEKL
```

```
HBA_HUMAN GSAQVKGHGKVKVADALTNVAHVV--D--DMPNALSALSOLHAEKL
++ ++++H+ KY + +A ++ +L+ L+++H+ K
LGB2_LUPLU NNPQLQAHGKVKVLPVYEAALQLQVTVGVVVDATLKNLGSVHVEKG
```

```
HBA_HUMAN GSAQVKGHGKVKVADALTNVAHVVDDMPNALSALS----LHAEKL
GS+ + G + +D L ++ H+ D+ A +AL D ++AH+
F11G11.2 GSGYLVGDSLTPVDELL--VAQHTADLLAANAALLDEFFQKAEQE
```

Bioinformatics lecture March 5, 2002



## Alignment scoring

- the 1st alignment: highly significant
- the 2nd: plausible
- the 3rd: spurious
- distinguish by *alignment score*
- similarities increase score
- mismatches decrease score substitution matrix
- gaps decrease score gap penalties

Bioinformatics lecture March 5, 2002



## Substitution matrices

- Substitution matrix weights replacement of one residue by another:
  - similar -> high score (positive)
  - different -> low score (negative)
- simplest is identity matrix (e.g. for nucleic acids)

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

Bioinformatics lecture March 5, 2002

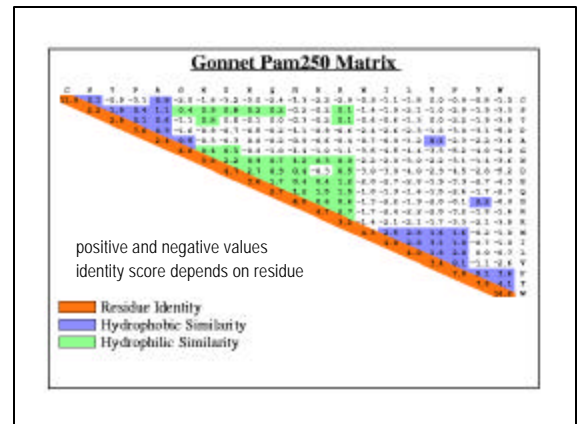


## Derivation of substitution matrices PAM matrices

73

- **PAM matrix series (PAM1 ... PAM250):**
  - derived from alignment of very similar sequences
  - PAM1 = mutation events that change 1% of AA
  - PAM2, PAM3, ... extrapolated by matrix multiplication
  - e.g.: PAM2 = PAM1 \* PAM1; PAM3 = PAM2 \* PAM1 etc
- **Problems with PAM matrices:**
  - incorrect modelling of long time substitutions, since
  - conservative mutations dominated by single nucleotide change
  - e.g.: L <-> I, L <-> V, Y <-> F
  - long time: any AA change

Bioinformatics lecture March 5, 2002



## BLOSUM matrices

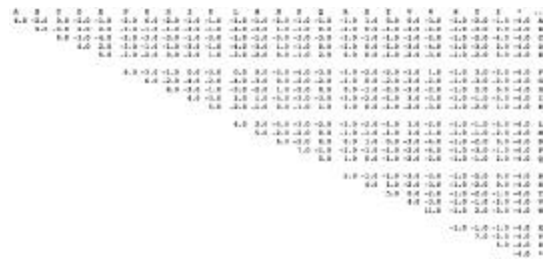
75

- **BLOSUM series (BLOSUM50, BLOSUM62, ...)**
- derived from alignments of distantly related sequence
- **BLOCKS database:**
  - ungapped multiple alignments of protein families at a given identity
- BLOSUM50 better for gapped alignments
- BLOSUM62 better for ungapped alignments

Bioinformatics lecture March 5, 2002



## Blosum62 substitution matrix



## Gap penalties

77

- significance of alignment:
  - depends critically on gap penalty
- need to adjust to given sequence
- gap penalties influenced by knowledge of structure etc
- simple rules when nothing is known (linear or affine)

Bioinformatics lecture March 5, 2002



## Gap penalties

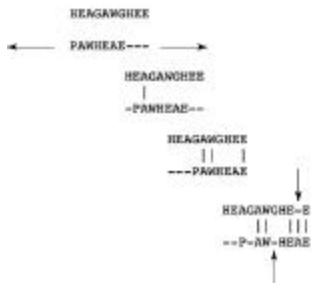
78

- **linear gap penalty:** one constant  $d$  for each insertion  $g$   
 $\mathcal{J}(g) = -g * d$  with  $g$  length of gap
- **affine gap penalty:**
  - (large) penalty  $d$  for opening of gap
  - (smaller) penalty  $e$  for extension of existing gap $\mathcal{J}(g) = -d - (g-1) * e$ , with  $g$  = length of gap
- example  $d = 10, e = 0.2$

Bioinformatics lecture March 5, 2002



## Alignment of two sequences



Bioinformatics lecture March 5, 2002



## Alignment algorithms

- maximize score:  
match as many positively scoring pairs as possible
- minimize cost:  
reduce number of mismatches and number of gaps
- possibilities to align 2 sequences of length  $n$ :

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

Bioinformatics lecture March 5, 2002



## Dynamic programming algorithm

- dynamic programming =  
build up optimal alignment  
using previous solutions  
for optimal alignments of subsequences

Bioinformatics lecture March 5, 2002



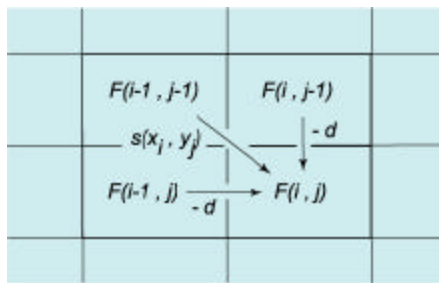
## Dynamic programming algorithm

- define a matrix  $F_{ij}$ :  
 $F_{ij}$  is the optimal alignment of  
subsequence  $A_1...i$  and  $B_1...j$
- iterative build up:  $F(0,0) = 0$
- define each element  $i,j$  from  
( $i-1,j$ ): gap in sequence A  
( $i,j-1$ ): gap in sequence B  
( $i-1,j-1$ ): alignment of  $A_i$  to  $B_j$

Bioinformatics lecture March 5, 2002



## Dynamic programming



Bioinformatics lecture March 5, 2002



## Scores from substitution matrix

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	5	0	5	-3	0	-2	-1	-1
W	-3	-3	-3	-3	-3	15	-3	-3	-3	-3
H	10	0	-2	-2	-2	-3	-2	10	0	0
E	0	6	-1	-3	-1	-3	-3	0	6	6
A	-2	-1	5	0	5	-3	0	-2	-1	-1
E	0	6	-1	-3	-1	-3	-3	0	6	6

Bioinformatics lecture March 5, 2002



### (1) Initialize boundaries

	H	E	A	G	A	W	G	H	E	E	
P	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
A	-8										
W	-16										
H	-24										
E	-32										
A	-40										
A	-48										
E	-56										

Bioinformatics lecture March 5, 2002



### (2) Fill matrix with minimum score sums..

	H	E	A	G	A	W	G	H	E	E	
P	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
A	-8	-2									
W	-16										
H	-24										
E	-32										
A	-40										
A	-48										
E	-56										

$0 + s(H,P) = 0 - 2 = -2$   
 $-8 + (-8) = -16$

Bioinformatics lecture March 5, 2002



### from top left corner

	H	E	A	G	A	W	G	H	E	E	
P	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
A	-8	-2	-9								
W	-16	-10	-3								
H	-24	-18	-11								
E	-32	-14	-8								
A	-40	-22	-8								
A	-48	-30	-16								
E	-56	-38	-24								

Bioinformatics lecture March 5, 2002



### Filled matrix: score in right bottom corner

	H	E	A	G	A	W	G	H	E	E	
P	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
A	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
W	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
H	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
E	-32	-14	-8	-13	-8	-9	-13	-7	-3	-11	-19
A	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

Bioinformatics lecture March 5, 2002



### (3) Backtracing gives alignment

	H	E	A	G	A	W	G	H	E	E	
P	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
A	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
W	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
H	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
E	-32	-14	-8	-13	-8	-9	-13	-7	-3	-11	-19
A	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

Bioinformatics lecture March 5, 2002



### Alternative optimum alignment

	H	E	A	G	A	W	G	H	E	E	
P	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
A	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
W	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
H	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
E	-32	-14	-8	-13	-8	-9	-13	-7	-3	-11	-19
A	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

Bioinformatics lecture March 5, 2002



## Alignment algorithms

91

- **global alignment (ends aligned)**
  - Needleman & Wunsch, 1970
- **local alignment (subsequences aligned)**
  - Smith & Waterman, 1981
- **searching for repetitions**
- **searching for overlap**

Bioinformatics lecture March 5, 2002



## Example output of GCG program bestfit

92

- alignment score depends on score matrix
- percent similarity - percent identity
- affine gap penalty favours grouping of gaps

Bioinformatics lecture March 5, 2002



```

Symbol comparison table: /usr/infobiogen/pack/gcg-9.1/gcgcore/dna/mxdata/kin
aa2.rap Compcheck: 6430

      Gap Weight: 12      Average Match: 2.912
      Length Weight: 4      Average Mismatch: -1.003

      Quality: 2432      Sequels: 662
      Ratio: 3.753      Gaps: 4
      Percent Similarity: 79.598      Percent Identity: 72.790

Match display thresholds for the alignment(s):
  | = 1000000
  + = 2
  . = 1

aaabab.seq x aaabab.seq May 6, 1990 11:58 ..
      2  EHWFFELCTTYYCOELEDKASSTFGGAMASSTELAGDYMELDDE 51
      3  |..|||:|||||:|||||:|||||:|||||:|||||:|||||:|||||
      4  EHWFFELCTTYYCOELEDKASSTFGGAMASSTELAGDYMELDDE 53
      53  EHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 161
      54  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      62  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      63  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      64  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      65  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      66  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      67  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      68  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      69  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      70  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      71  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      72  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      73  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      74  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      75  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      76  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      77  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      78  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      79  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      80  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      81  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      82  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      83  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      84  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      85  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      86  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      87  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      88  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      89  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      90  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      91  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      92  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      93  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      94  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      95  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      96  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      97  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      98  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
      99  *|||||:|*:*:*|||*||| -|*||*|*|*|*|*|*|*|*|*|*|*|
      100  GHWGKTSKAGKQVTFQQTVEGVVIGISAKSLKSTYDQFVYVYDHR 163
  
```

## Database searches: FASTA and BLAST

94

- Full Smith-Waterman search expensive ( $O(mn)$ )
- database contains > 100 million residues
- heuristic programs concentrate on important regions
- evaluate few cell in the dynamic programming matrix

Bioinformatics lecture March 5, 2002

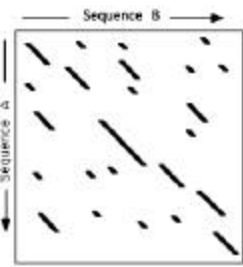


## FASTA

95

- multi-step approach to find high-scoring alignments
- (1) exact short word matches
- (2) maximal scoring ungapped extensions
- (3) identify gapped alignments

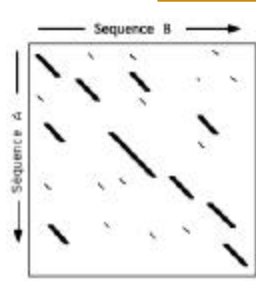
Bioinformatics lecture March 5, 2002



- lookup table to find all identically matching words
- length  $ktup$
- $ktup = 1,2$  for proteins
- $ktup = 4-6$  for DNA

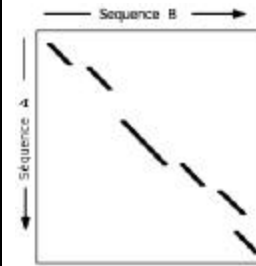
Bioinformatics lecture March 5, 2002





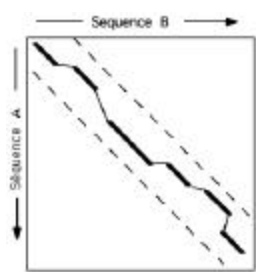
- Scoring the words with the substitution matrix

Bioinformatics lecture March 5, 2002



- extend exact word matches to find maximal scoring ungapped regions

Bioinformatics lecture March 5, 2002



- join ungapped regions in one gapped region
- highest scoring candidate matches are realigned in a narrow band around match

Bioinformatics lecture March 5, 2002



## BLAST

- multi-step approach to find high-scoring alignments
- (1) list words of fixed length (3AA) expected to give score larger than threshold
- (2) for every word, search database and extend ungapped alignment in both directions
- (3) new versions of BLAST allow gaps

Bioinformatics lecture March 5, 2002



## BLAST program suite

- various versions:
  - blastn**: nucleotide sequences
  - blastp**: protein sequences
  - tblastn**: protein query - translated database
  - blastx**: nucleotide query - protein database
  - tblastx**: nucleotide query - translated database

Bioinformatics lecture March 5, 2002



<http://www.ncbi.nlm.nih.gov/BLAST>



# BLAST

BLAST info

What's NEW in BLAST®  
September 26, 2001: New gapps are now available for Mouse, Rat and Fugu genomes in the "Genomic BLAST pages" section.

Results from the Protein Sequence Information Survey  
NCBI's Protein Sequence Information Survey Results are [here](#). Thank you for participating.

### Nucleotide BLAST

- Standard nucleotide-nucleotide BLAST (blastn)
- MEGABLAST
- Search for short nearly exact matches

2

URL API documentation

### Protein BLAST

- Standard protein-protein BLAST (blastp)
- PSI and PSI-BLAST
- Search for short nearly exact matches

2

FTP

### Translated BLAST Searches

- Nucleotide query - Protein db (blastx)
- Protein query - Translated db (tblastn)
- Nucleotide query - Translated db (tblastx)

2

Credits

.....

## Multiple sequence alignment and sequence profiles

103

- Scoring a multiple sequence alignment
- An alignment algorithm: CLUSTALW
- Sequence profiles and profile searches

Bioinformatics lecture March 5, 2002



## Multiple sequence alignment

104

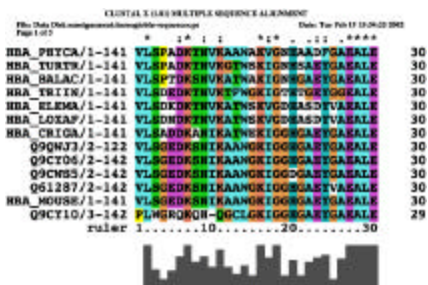
- compare set of sequences
- align homologous residues in columns
- homologous residues:
  - evolutionary: diverge from common ancestral residue
  - structurally: occupy similar position in space
- generally impossible to get single "correct" alignment
- focus on key residues and align them in columns

Bioinformatics lecture March 5, 2002

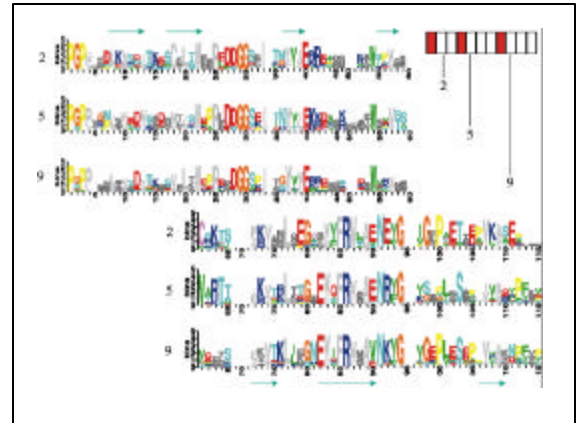


## Example: part of haemoglobin alignment

105



Bioinformatics lecture March 5, 2002



## Scoring multiple sequence alignment

107

- take into account:
  - (1) some positions more conserved than others
  - (2) sequences are not independent but related in a phylogenetic tree
- approximation: assume columns of alignment are statistically independent
- total score of alignment is sum of column scores
- each column score is a sum of all sequence pairs

Bioinformatics lecture March 5, 2002



## Multiple sequence alignment algorithms

108

- multidimensional dynamic programming
  - very expensive, only possible for few sequences
- progressive alignment methods
  - construct a series of pair-wise alignments

Bioinformatics lecture March 5, 2002



## CLUSTALW and CLUSTALX

109

- align all sequence pairs by dynamic programming
- convert alignment into evolutionary distances
- construct a "guide tree"
- align nodes of the tree in order of decreasing similarity
  - sequence-sequence
  - sequence-profile
  - profile-profile alignment

Bioinformatics lecture March 5, 2002



## Guide tree

110



- Guide tree is a "quick and dirty" phylogenetic tree
- Clustal alignment starts at the right (the leaves)
- progresses to the left
- aligned sequences:
  - sequence profile

Bioinformatics lecture March 5, 2002



## CLUSTAL

111

- other important features:
- sequences are weighted to compensate for bias
- substitution matrix depending on expected similarity:
- similar sequences with "hard" matrices (BLOSUM80)
- distant sequences with "soft" matrices (BLOSUM50)
- position specific gap open penalties

Bioinformatics lecture March 5, 2002



## Sequence profiles

112

- multiple sequence alignment -> sequence profile
- evolutionary relationship
- "sequence-specific substitution matrix"
- very sensitive database searches

Bioinformatics lecture March 5, 2002



## Sequence profile

113



Bioinformatics lecture March 5, 2002



## Profile searches

114

- "by hand":
  - database search (Smith-Waterman)
  - multiple sequence alignment
  - calculation of profile
  - profile database search
- possible at <http://eta.embl-heidelberg.de:8000>
- less sensitive but much easier: psi-blast at NCBI

Bioinformatics lecture March 5, 2002

