Improved tools for biological sequence comparison

Authors: William R. Pearson, David J. Lipman

Talk given by: Niclas Nordholt

Table of contents

The authors

- William R. Pearson
  - Ph.D.,California Institute of Technology, 1977
  - Professor of Biochemistry at the University of Virginia

The authors

- David J. Lipman
  - B.A. (Honors) Brown University, Rhode Island, 1976
  - M.D. University of New York at Buffalo, 1980
  - Since 1989 director of NCBI, home of PubMed and GenBank

FASTP

- Published in 1985
- Stands for "FAST-Protein"
- A heuristic program for protein sequence similarity searching
- Predecessor of the well known FASTA

FASTA

- Published in 1988
- A more sensitive derivative of FASTP
- FASTA stands for "FAST-All" because it works on any alphabet
  - Protein:Protein, DNA:DNA, Protein:DNA
  - But also for any alphabet with arbitrary match/mismatch scoring values
  - All scoring parameters can be interchanged without changing the program

FASTA

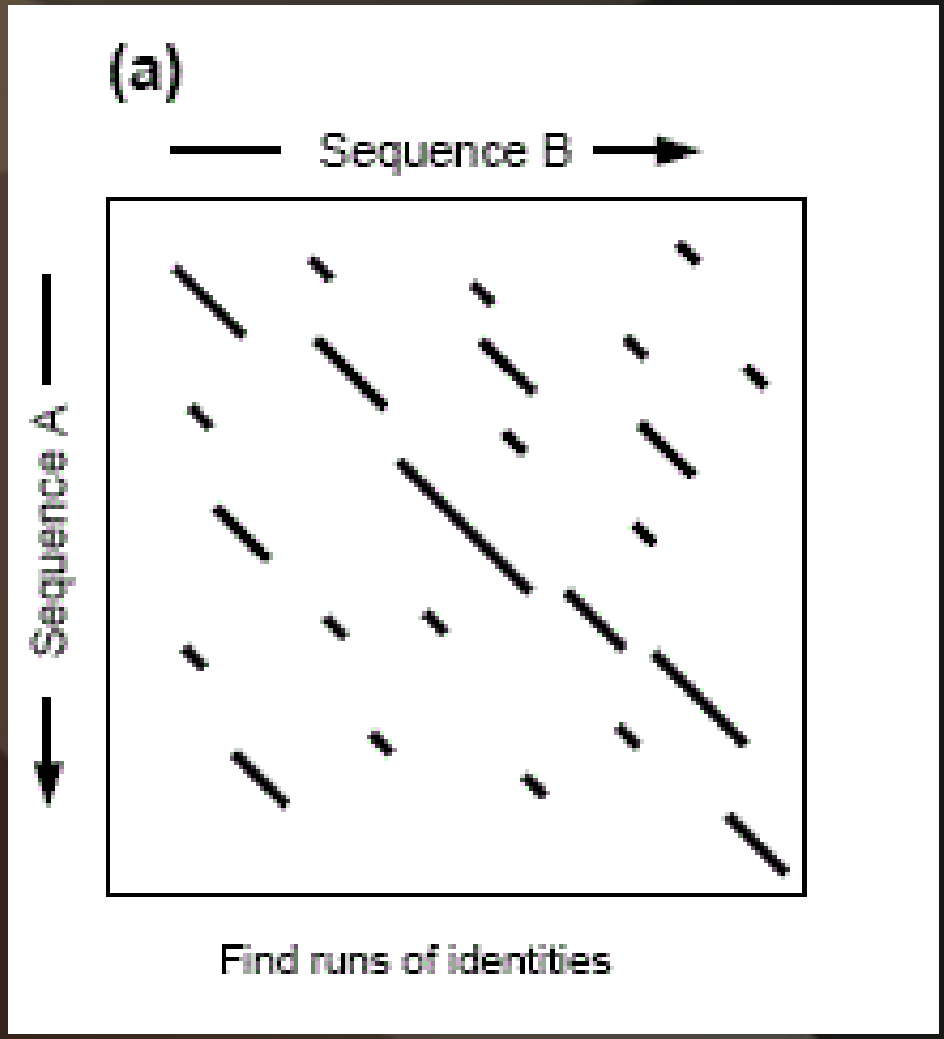
The FASTA-Algorithm can be divided into 4 steps:

FASTA – Step 1

- Locate all identities or groups of identities between two sequences by using a lookup table
  - The lookup table has to be pre-processed
  - The *ktup* parameter determines the number of consecutive identities required for a match
  - e.g. if *ktup*=4, identities with less than a run of 4 matches are not considered
  - The lower the *ktup* parameter, the higher the sensivity BUT the lower the selectivity (unrelated sequences!)

FASTA – Step 1

- Recommended *ktup* values:
  - DNA: 4 to 6
  - Protein: 2
  - Short sequences: 1
- The best 10 regions are saved

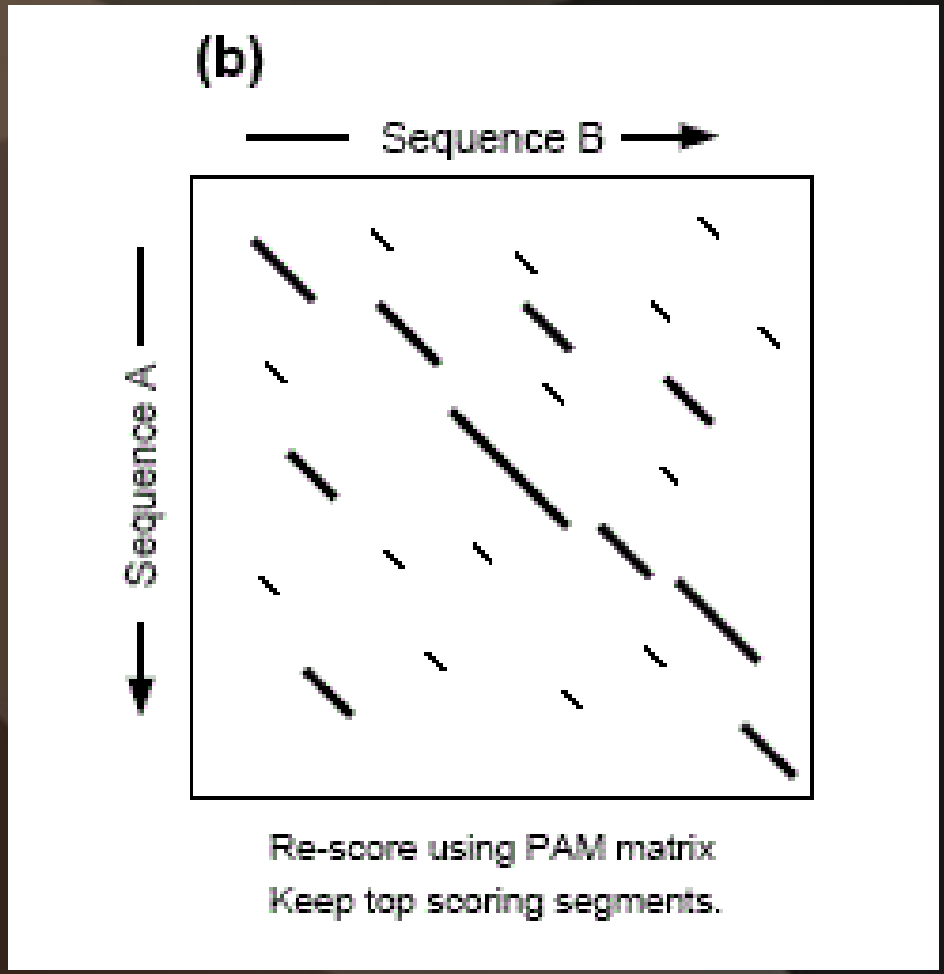


FASTA – Step 2

- Rescore these 10 regions using a scoring matrix (e.g. PAM250)
  - Conservative replacements and runs of identities < *ktup* are allowed to contribute to similarity score
  - For each of the best diagonal regions rescanned with the scoring matrix, a subregion with the maximal score is identified (*initial region*)
  - The best initial region is referred to as *init1*

FASTA – Step 2

- The best initial regions are shown as bold diagonals

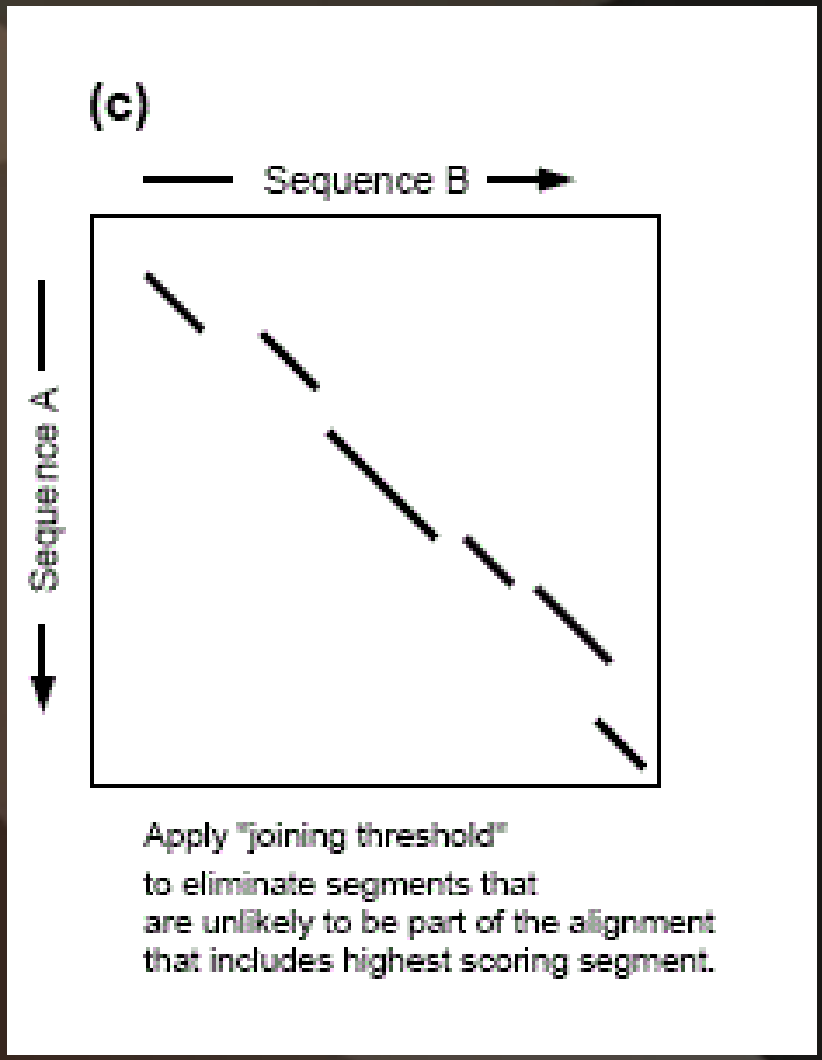


FASTA – Step 3

- Check whether the initial regions can be joined together, forming an optimal alignment with gaps
  - Increases sensitivity thus decreases selectivity
  - Selectivity is preserved by only using initial regions above a joining threshold *t*
  - Calculate a similarity score that is the sum of the joined regions and the gap penalties
  - These initial scores are used to rank the library sequences

FASTA – Step 3

- The joinig threshold *t* has to be chosen carefully to preserve selectivity!
- e.g. for a query sequence of 200 residues and *ktup*=2 use *t*=28

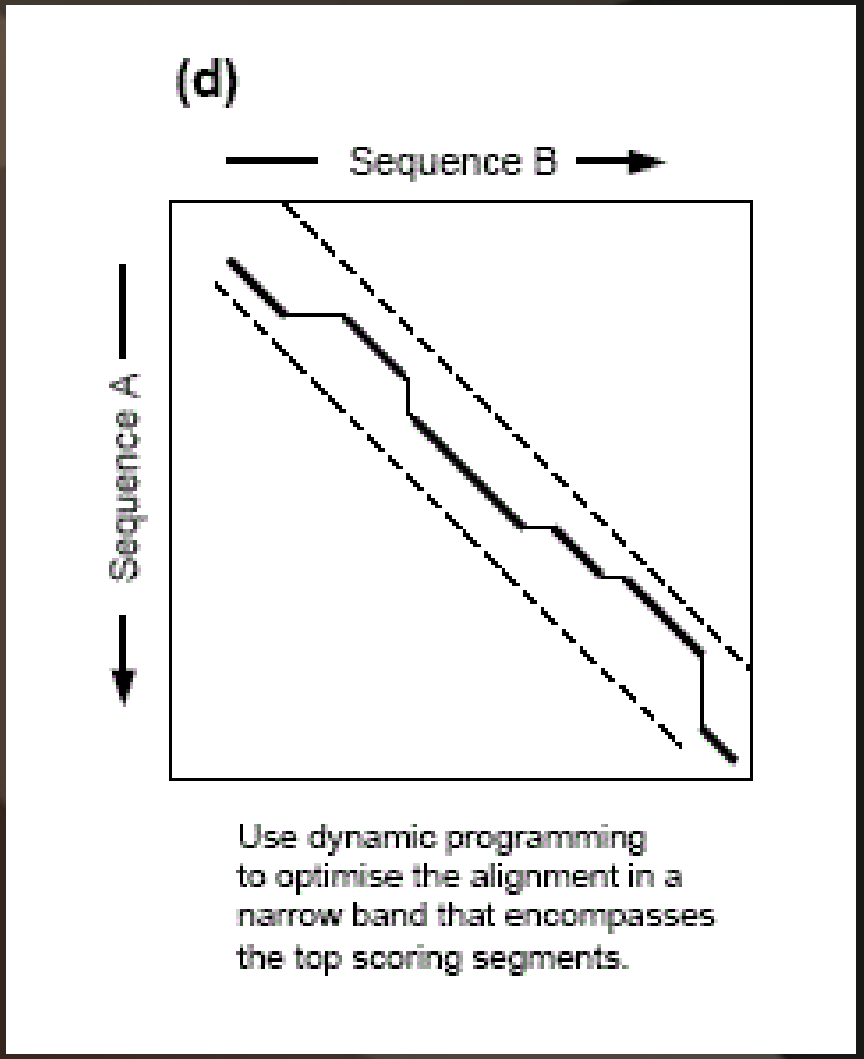


FASTA – Step 4

- Align the highest scoring library sequences using a modification of the Smith-Waterman algorithm
  - Considers all possible alignments of the query and library sequences that lie within a band centered around the best initial region (*init1*)
  - The result of this alignment is reported as the optimized score

FASTA – Step 4

- The dotted lines denote the bounds of the optimal alignment



FASTP vs FASTA

- In step 3 FASTP would only use *init1* whereas FASTA tries to join several initial regions
- =>higher sensitivity/scores for related sequences

Table 1. FASTA and FASTP initial scores of the T-cell receptor (RWMSAV) versus the NBRF data base

NBRF code	Sequence	Initial score	
		FASTA	FASTP
RWHUAV	T-cell receptor α chain	155	98
K1HURE	Ig κ chain V-I region	127	111
KVMS50	Ig κ chain V region	149	62
KVMSM6	Ig κ chain precursor V regions	141	64
KVRB29	Ig κ chain V region	126	54
L3HUSH	Ig λ chain V-III region	90	47
KVMS41	Ig κ chain precursor V region	87	87
RWMSBV	T-cell receptor β-chain precursor	94	94
RWHUVY	T-cell receptor β-chain precursor	91	59
RWHUGV	T-cell receptor γ-chain precursor	87	61
RWHUT4	T-cell surface glycoprotein T4	86	63
RWMSVB	T-cell receptor γ-chain precursor	71	41
HVMS44	Ig heavy-chain V region	67	36
G1HUDW	Ig heavy-chain V-II region	62	35

The average FASTP score = 26.1 ± 6.8 (mean ± SD). The average FASTA score = 26.2 ± 7.2 (mean ± SD). The mean and SD were computed excluding scores >54. V, Variable.

TFASTA

- Variaton of FASTA
- The "T" stands for "translated"
- Amino acid sequence comparisons are more sensitive than DNA sequence comparisons
  - employment of scoring matrices!
- Protein sequence is compared against DNA library
  - Translation of all six reading frames "on-the-fly"

TFASTA

Table 2. DNA data base search of rat transforming growth factor (RATTGFA) versus mammalian sequences

GenBank locus	Sequence	Score	
		Initial	Optimized
HUMTFGAM	Human TGF mRNA	1336	1618
HUMTGFA2	Human TGF gene (exon 2)	354	366
HUMTGFA1	Human TGF gene (5′ end)	224	381
MUSRGEB3	Mouse 18S–5.8S–28S rRNA gene	140	107
MUSRGE52	Mouse 18S–5.8S–28S rRNA gene	140	107
MUSMHDD	MHC class I H-2D	122	78
HUMMETIF1	Metallothionein (MT)I_F gene	116	92
MUSRGLP	45S rRNA (5′ end)	115	83
HUMPS2	pS2 mRNA	105	106
MUSC1AI1	α-1 type I procollagen	86	89

The 10 sequences having the highest initial scores are given. TGF, transforming growth factor; MHC, major histocompatibility complex.

Table 3. Translated DNA data base search of rat transforming growth factor (RATTGFA) versus mammalian sequences

GenBank locus	Sequence	Frame	Score	
			Initial	Optimized
RATTGFA	Rat TGF type α	1	816	816
HUMTGFAM	Human TGF mRNA	2	671	770
HUMTGFA2	Human TGF gene	1	204	205
MUSEGF	Mouse EGF mRNA	3	93	129
MUSMHAB3	Mouse MHC class II H2-IA_β	1	91	58
MUSIGCD17	Mouse Ig germ-line DJC region	3′	85	48
HUMESTR	Human estrogen receptor	3	83	65
RATINSI	Rat insulin 1 (*Ins-1*) gene	2	81	63
MUSTHYS1	Mouse thymidylate synthase	2	80	63
HUMPNU3	Human purine nucleoside phosphorylase	1′	80	52

The 10 sequences having the highest initial scores are given. TGF, transforming growth factor; EGF, epidermal growth factor; D, diversity; J, joining; C, constant; MHC, major histocompatibility complex.

LFASTA

- A program for detecting local similarities
- Uses the same first two steps as FASTA
- Can identify multiple alignments between smaller portions of two sequences
- Saves ALL initial regions above a threshold
- Computes a local alignment for each initial region
- Thus repeats or duplications can be found

LFASTA

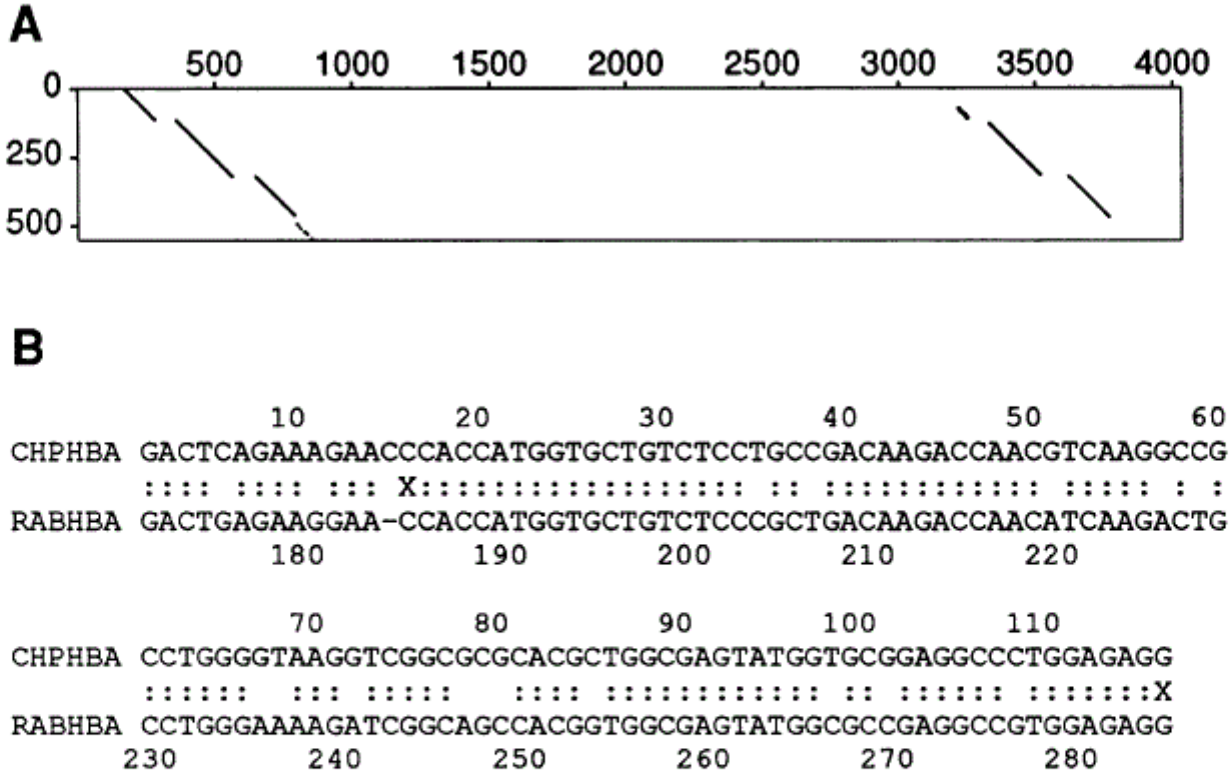


FIG. 2. Local comparison of an α-globin mRNA sequence with an α-globin gene cluster. An ape α_1-globin mRNA sequence (GenBank sequence CHPHBA1M) was compared with a rabbit α-globin gene sequence (RABHBAPT) containing a second pseudo-θ-globin gene using the LFASTA program. (*A*) A plot of the homologous regions shared by the two sequences. (*B*) One of the alignments between the mRNA sequence and the rabbit α-globin gene (nucleotides 171–855). Three other alignments between the mRNA sequence and the α-globin gene and three alignments between the pseudo-θ-globin gene (nucleotides 3200–3770) were calculated but are not shown. There is 84.3% identity in the 115 nucleotide overlap. The initial region and optimized scores using LFASTA are 284 and 304, respectively. X denotes the ends of the initial region found by LFASTA.

LFASTA

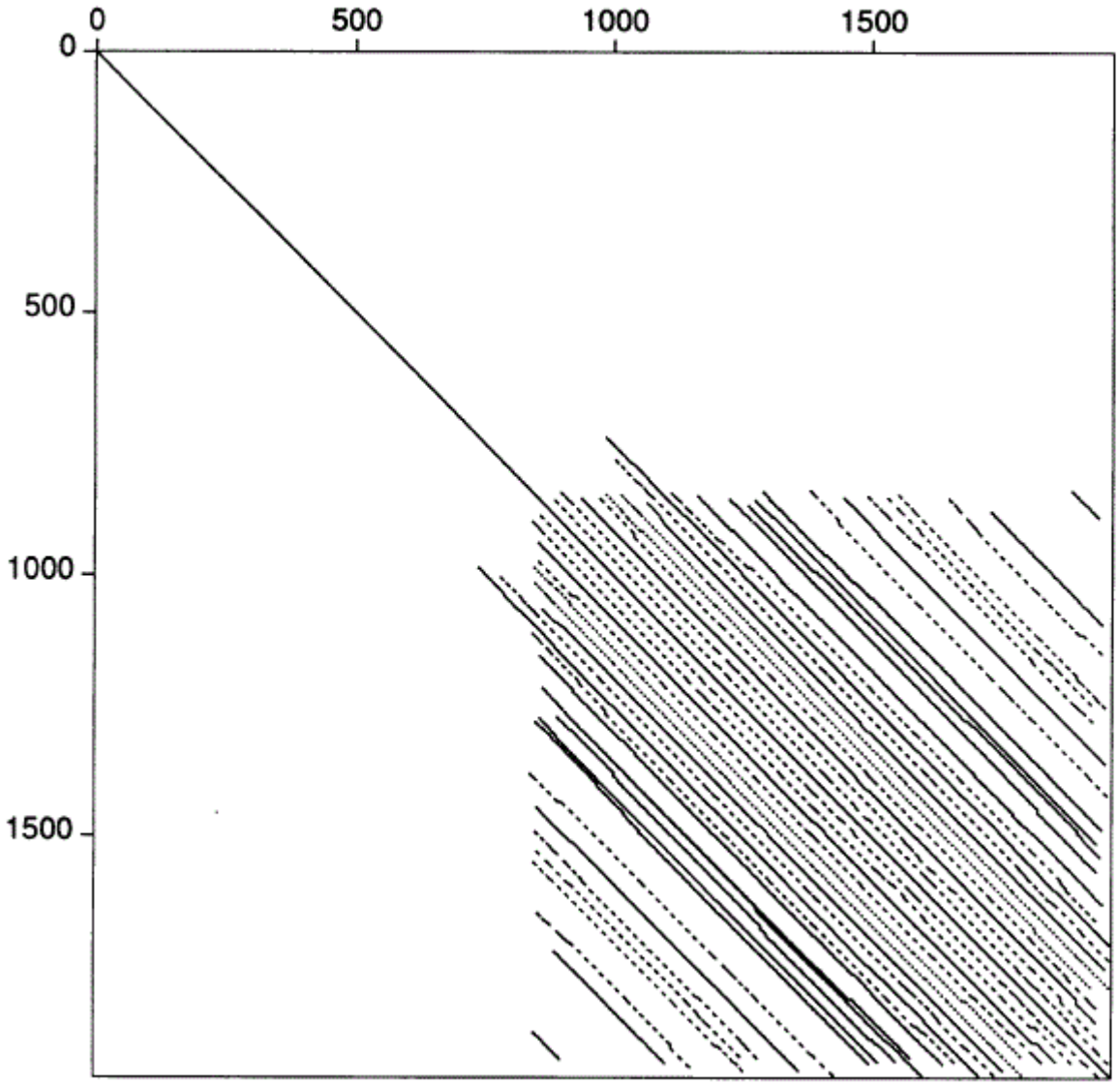


FIG. 3. Repeated structure in the myosin heavy chain. LFASTA was used to compare the *Caenorhabditis elegans* myosin heavy chain protein sequence (NBRF code MWKW) with itself using the PAM250 scoring matrix. The solid, dashed, and dotted lines denote decreasing similarity scores. The solid lines had initial region scores greater than 80 and optimized local scores greater than 150; the longer dashed lines had initial region and optimized local scores greater than 65 and 120, respectively, and the shorter dashed lines had initial region and optimized local scores greater than 50 and 100, respectively. Homologous regions with lower scores are plotted with dots.

RDF2

- Searching a library with any scoring program will find a highest score, regardless of whether there is a biological relation between these two sequences or not
- Therefore the statistical significance has to be evaluated
- RDF2 does a Monte Carlo shuffle analysis on the found sequences

RDF2

- RDF2 calculates 3 scores for each shuffled sequence
  - Best single initial region
  - Joined initial regions
  - Optimized diagonal
- Allows employment of a scoring matrix defined by the user
- There is a global and a local shuffle routine

RDF2

- Global shuffle routine: taking each residue of a sequence an placing it randomly along the length of the new sequence
- Local shuffle routine: permutation of small blocks of 10 to 20 residues
  - Order of the sequence is destroyed, local composition is not

Conclusion

- There is always a trade off between sensitivity and selectivity
- While trying to find the most likely set of mutations any tool for sequence similarity analysis must contain an implicit model of molecular evolution
  - Scoring rules must fit evolution
- Careful evaluation is important

References

- Improved tools for biological sequence comparison. Pearson WR, Lipman DJ. Department of Biochemistry, University of Virginia, Charlottesville 22908
- http://faculty.virginia.edu/wrpearson/
- http://www.ncbi.nlm.nih.gov/CBBresearch/Lipman/
- http://en.wikipedia.org/wiki/FASTA
- http://www-bimas.cit.nih.gov/fastainfo/fasta_algo