Basic Local Alignment Search Tool (BLAST)

Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, David Lipman

Journal of Molecular Biology 1990

Presented for MBB752
Brian Dunican
January 28, 2009

The Problem

- Dynamic programming algorithms take too long when applied to large databases
- Also these algorithms maximize similarity
  - Insertions
  - Deletions
  - Replacements
- Example Needleman and Wunsch

BLAST

- Similarity measure is based on well defined mutation scores (PAM120)
  - Optimization of this measure approximates the results of dynamic programming algorithm

- "Detect weak but biologically significant sequence similarities, and is more than an order of magnitude faster"

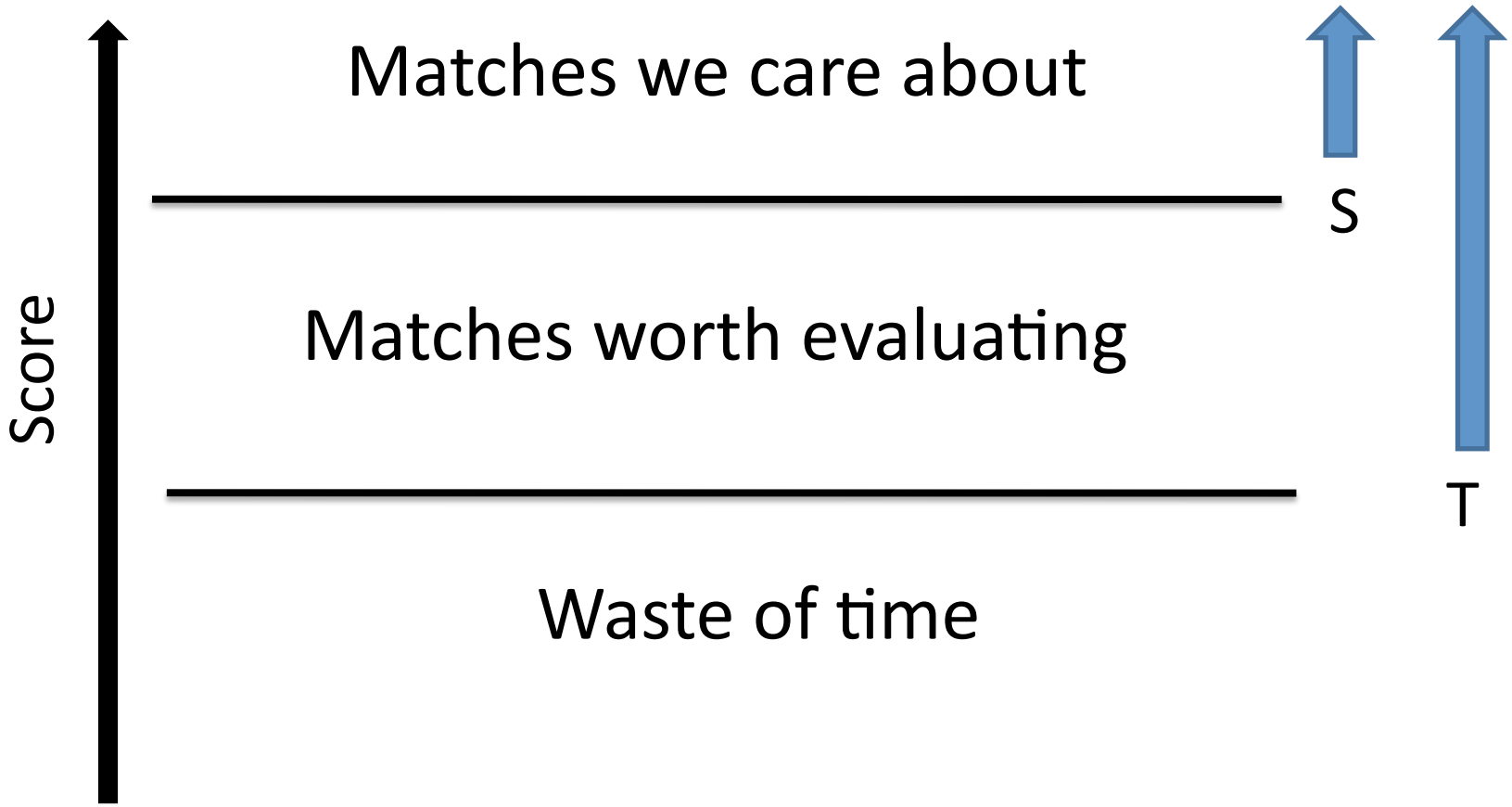
Methods

- Matrix of Similarity
  - Proteins: similar +, dissimilar – (PAM120)
  - DNA: identities +5, mismatches -4
- Units of Maximal Segment Pair (MSP)
  - Score can not be increased by shortening or lengthening the segment pair
- BLAST seeks the highest MSP score (>=S)

Method

- Computationally intensive to scan the database for all w-meres in search of S
- Define a threshold value, T , as the lower bound for further analysis.
- Database is searched for all words (w-meres) which can equal T.
  - Two steps
- From matches with score >= T, dynamic programming is used to determine MSP
  - Traces out from hit to maximize score

Method


Score
Matches we care about
S
Matches worth evaluating
T
Waste of time

Parameters

- The chance exists the a random sequence will exceed the S score

$$1 - e^{-y}, \qquad (1)$$

$$y = Kmn\ e^{-\lambda S}.$$

M and N are the length of the compared strings

S is the arbitrary score

K and lambda are coefficients

Parameters

- Used equation (1) to determine w and T parameters

w	T	Probability of a hit $\times 10^5$	Implied % of MSPs missed by BLAST when S equals						
			45	50	55	60	65	70	75
3	11	253	1	1	0	0	0	0	0
	12	147	4	3	2	1	1	0	0
	13	83	11	8	6	4	3	2	2
	14	48	20	16	12	10	8	6	5
	15	26	33	28	23	20	17	14	12
	16	14	46	41	36	32	29	26	23
	17	7	59	55	51	47	43	40	37
	18	4	70	67	63	60	57	54	51
4	13	127	2	1	1	0	0	0	0
	14	78	5	3	2	1	1	0	0
	15	47	10	7	5	4	3	2	1
	16	28	18	14	11	8	6	5	4
	17	16	28	23	19	16	13	11	9
	18	9	40	35	30	26	22	19	17
	19	5	51	46	41	37	33	30	27
	20	3	62	57	53	49	45	41	38
5	15	64	3	2	1	1	0	0	0
	16	40	6	4	3	2	1	1	0
	17	25	12	9	6	4	3	2	2
	18	15	20	15	12	9	7	5	4
	19	9	29	23	19	15	13	10	8
	20	5	38	32	28	23	20	17	14
	21	3	48	42	37	32	29	25	22
	22	2	57	52	47	42	38	35	31
Expected no. of random MSPs			50	9	2	0·3	0·06	0·01	0·002

Time Constraints

- Compile list of words that can score T from query
- Scan database for matches to T-scoring words
- Extend all hits to seek MSPs higher than S

- Increasing w decreases time spent on step 3.
- Make w too high and step one is limiting factor

Performance

- Against real proteins
  - Woolly monkey myoglobin (w=4, t=17)
    - Actual: missed 43 MSPs with a 50>S>80
    - Expected: miss 24 of 178
  - Mouse immunoglobulin precursor V region
    - Actual: missed 2 with a 45>S>65
    - Expected: miss 8 of 33
- Lost out in the monkey due to the uniform pattern of conservation
- Expect that on average Blast will outperform the random model.

Notes on Speed

- Proteins: 500,00 residues/s
- DNA: 2,000,000 bases/s