Sequence Comparison
Sequence Comparison

Legend: PhaC - PhaC1, PhaC2 USM 4.65  a/b Cons - α/β hydrolase fold consensus sequence
Why Compare Sequences?

“The first fact of biological sequence analysis – In biomolecular sequences (DNA, RNA, or amino acid sequences), high sequence similarity usually implies significant functional or structural similarity.”

How to Measure

• The most often used distance on strings in computer science is Hamming distance.

  AGTTTAATCA
  ||| ||| |
  AGTATAACGA

• There are other mutations during evolution
  – Substitution ACAGT → ACGGT
  – Insertion/deletion (indel) ACAGT → ACGT
  – Inversion ACA......GT → AG......ACT
  – Translocation AC......AG...TAA → AG...TC......AAA
  – Duplication AC...A...... → AC...A......C...A
  – ......

• There are many formulations of sequence comparison and many algorithms by considering different combinations of these mutations.

The Most Popular Measure

• Edit operations: substitution, insertion, deletion.

• Edit distance: Minimum number of edit operations to convert one sequence to another.

• Example: A text editor.
Edit Distance vs. Alignment

GAGTTTAATAC
   |||   |||   |
TAGTATAA–AG

• This is an alignment with score:
  7*match + 3*mismatch + 1*indel

• Under correct score scheme, edit distance is equivalent to sequence alignment.

• Alignment is easier to show.
Levenshtein

• Edit distance or Levenshtein distance.
• Vladimir Iosifovich Levenshtein (Russian: Влади́мир Ио́сифович Левенштей́н; born 1935) is a Russian scientist who has done research in information theory, error-correcting codes, and combinatorial design.
• In error correction coding. Not computational biology.

IEEE Richard W. Hamming Medal Recipient 2006
Needleman-Wunsch

• Two western researchers independently proposed this distance in computational biology.


• Cited by 6758.
Longest Common Subsequence

• Needleman-Wunsch actually used the term “maximum match” that is essentially the longest common subsequence.

• Let \( M(i,j) \) be the LCS of \( S[1..i] \) and \( T[1..j] \). The algorithm is essentially the following:

\[
M(i, j) = \max \begin{cases} 
M(i-1, j-1) + \delta(S[i], T[j]) \\
M(i-1, j) \\
M(i, j-1)
\end{cases}
\]

• The statistical significance is also briefly discussed.
  – This is important!
Backtracking

$$M(i, j) = \max \begin{cases} M(i-1, j-1) + \delta(S[i], T[j]) \\ M(i-1, j) \\ M(i, j-1) \end{cases}$$
In Terms of Edit Distance

• \( D(i,j) := \) edit distance between \( S[1..i] \) and \( T[1..j] \).

• Then

\[
D(i, j) = \min \begin{cases} 
D(i - 1, j - 1) + \chi(S[i], T[j]) \\
D(i - 1, j) + 1 \\
D(i, j - 1) + 1
\end{cases}
\]

• Time Complexity: \( O(mn) \)
In Terms of Alignment

**Edit distance**

\[ D(i, j) = \min \begin{cases} 
D(i - 1, j - 1) + \chi(S[i], T[j]) \\
D(i - 1, j) + 1 \\
D(i, j - 1) + 1 
\end{cases} \]

**Alignment**

\[ M(i, j) = \max \begin{cases} 
M(i - 1, j - 1) + score(S[i], T[j]) \\
M(i - 1, j) + indel \\
M(i, j - 1) + indel 
\end{cases} \]

LCS, Edit Distance, Alignments are closely related.
Linear Space

• 20 bits addressing space = 1MB memory.
• 16 Bits linear space = 64KB.
• 640KB limit for DOS programming.
• Special technique needed to use > 640KB.

<table>
<thead>
<tr>
<th>Generation</th>
<th>First introduced</th>
<th>Prominent consumer CPU brands</th>
<th>Linear/physical address space</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1978</td>
<td>Intel 8086, Intel 8088 and clones</td>
<td>16-bit / 20-bit (segmented)</td>
</tr>
<tr>
<td>2</td>
<td>1982</td>
<td>Intel 8086 and clones, NEC V20/V30</td>
<td>16-bit (30-bit virtual) / 24-bit (segmented)</td>
</tr>
<tr>
<td>3 (IA-32)</td>
<td>1985</td>
<td>Intel 80386 and clones, AMD Am386</td>
<td></td>
</tr>
<tr>
<td>4 (FPU)</td>
<td>1989</td>
<td>Intel486 and clones, AMD Am486/Am5x86</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1993</td>
<td>Pentium, Pentium MMX, Cyrix 5x86, Rise mP6</td>
<td></td>
</tr>
</tbody>
</table>

Cited by 847.
First Trick

• Compute column by column and only keep two recent columns.
• $O(m)$ space.
• Drawback: can only know the score of one (or two) columns.
The Algorithm

• Divide and Conquer.
• In $O(mn)$ time and $O(m + n)$ space, find $k$ such that $LCS(S[1..k], T[1..\frac{n}{2}]) + LCS(S[k + 1..m], T[\frac{n}{2} + 1..n])$ is maximized.
• Then recursively solve the two smaller problems.

$$T(m, n) = mn + T\left(k, \frac{n}{2}\right) + T\left(m - k, \frac{n}{2}\right) = O(mn)$$
How to Find \( k \)?

\[
\text{LCS}(S[1..k], T[1..\frac{n}{2}]) \quad \text{and} \quad \text{LCS}(S[k+1..m], T[\frac{n}{2} + 1..n])
\]
Gaps Are not Welcomed

- Think of deleting a consecutive row of letters during editing. It’s “cheaper” than scattered deleting.
- Evolution is somewhat similar.
- So we want to minimum number of gaps (not necessarily number of indels)

```
AGT--AATCA
  ||  ||  
AGTTCAACGA
```
David Sankoff

- David Sankoff first proposed a way to restrict gap.
  - Longest common subsequence.
  - Restrict the number of gaps to be $\leq K$.

*Proc. Nat. Acad. Sci. USA*
Vol. 69, No. 1, pp. 4–6, January 1972

**Matching Sequences under Deletion/Insertion Constraints**  
(algorithm/genetic homology/nucleotide sequence/amino-acid sequence)

DAVID SANKOFF

Centre de recherches mathématiques, Université de Montréal, c.p. 6128, Montréal 101

*Communicated by Mark Kac, October 18, 1971*

Cited 245 times.
What Is a Gap in LCS

AGT--AATCA

AGTTCAACGA

ABSTRACT Given two finite sequences, we wish to find the longest common subsequences satisfying certain deletion/insertion constraints. Consider two successive terms in the desired subsequence. The distance between their positions must be the same in the two original sequences for all but a limited number of such pairs of successive terms. Needleman and Wunsch gave an algorithm
LCS with No More than K Gaps

- $V_q(i, j) = \text{LCS of } S[1..i] \text{ to } T[1..j] \text{ with at most } q \text{ gaps, AND last match is \"on the same diagonal\" of } i \text{ vs. } j.$

- If $S[i] = T[j],$

$$V_q(i, j) = 1 + \max \begin{cases} V_q(i - 1, j - 1) \\ \max_{k > 0} V_{q-1}(i - k, j - 1) \\ \max_{k > 0} V_{q-1}(i - 1, j - k) \end{cases}$$

- Else,

$$V_q(i, j) = V_q(i - 1, j - 1)$$

- A straightforward implementation takes $O(n^3K).$
More Efficient Algorithm

• Sankoff introduced another matrix $W_q(i,j) := \text{LCS of } S[1..i] \text{ to } T[1..j] \text{ with at most } q \text{ gaps.}$
  
  — Same as $V$, except that no “same diagonal” restriction.
  
  $W_q(i, j) = \max\{W_q(i - 1, j), W_q(i, j - 1), V_q(i, j)\}$

• Now $V$ and $W$ can be computed simultaneously

  ```latex
  \begin{align*}
  V_q(i, j) &= \delta(S[i], T[j]) + \max\{V_q(i - 1, j - 1), W_{q-1}(i - 1, j - 1)\} \\
  W_q(i, j) &= \max\{W_q(i - 1, j), W_q(i, j - 1), V_q(i, j)\}
  \end{align*}
  ```

• $O(n^2K)$ time.

• This is like magic...
The advantage of this method over that of Needleman and Wunsch, as well as others that have been published, is that it does not depend on any arbitrarily imposed numerical criteria such as cell weights, but on a genetically meaningful criterion, the DI index. Further, this algorithm can be used as the basis for statistically testing hypotheses not only about the similarity of two sequences, but also about the number of deletions and insertions separating them (6).

I thank R. J. Cedergren for introducing me to this problem.


A few notes about this particular research
1. Small number of citations.
2. Statistics
3. Use of abstract notations (V,W).
Gap Penalty

• Instead of limiting # of gaps, Waterman, Smith and Beyer proposed to penalize them.
• Gap penalty: $g(k)$ is the cost of $k$-consecutive deletions.
• Normally $g(k) < k \times g(1)$. 

Some Biological Sequence Metrics*

M. S. Waterman
Idaho State University, Pocatello, Idaho 83209

T. F. Smith
Northern Michigan University, Marquette, Michigan 49855

AND

W. A. Beyer
Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545*
When There’s a Gap Penalty

\[
D(i, j) = \max \left\{ D(i - 1, j - 1) + \delta(S[i], T[j]) \right. \\
\left. \quad \max_k D(i - k, j) + g(k) \right. \\
\left. \quad \max_k D(i, j - k) + g(k) \right. 
\]

Time complexity: \( O(mnL) \)

L is maximum length of a gap.
Local Alignment

• Biologists do not necessarily know the start and end of the comparison.
  – E.g. Proteins have domains.

• To find a substring of S and a substring of T, so that their alignment score is maximized.

• Question: Why not “local edit distance”? 
Smith-Waterman

This simple change from Needleman-Wunsch made a big difference.
Affine Gap Penalty

• Recall that when Waterman et al. introduced the gap penalty $g(k)$, time was cubic.

• Osamu Gotoh noticed that when the gap penalty is affine: $g(k)=a+b^*(k-1)$, the time can be quadratic.


A: gap open
B: gap extension

An Improved Algorithm for Matching Biological Sequences

The algorithm of Waterman et al. (1976) for matching biological sequences was modified under some limitations to be accomplished in essentially $MN$ steps, instead of the $M^2N$ steps necessary in the original algorithm. The limitations do not seriously reduce the generality of the original method, and the present method is available for most practical uses. The algorithm can be executed on a small computer with a limited capacity of core memory.
Affine Gap Penalty

• In a paper 1986, Altschul gave an example where Gotoh’s algorithm may fail due to a glitch, and proposed a new algorithm.
Affine Gap Penalty

$M_0[i,j]$: best alignment score of $S[1..i]$, $T[1..j]$, such that $S[i]$ aligns with $T[j]$.  
$M_1$: $T[j]$ aligns with -  
$M_2$: $S[i]$ aligns with -  

\[
\begin{align*}
\text{ATAGG} & & \text{ATAGG} & & \text{ATAGGC} \\
\text{ATGG} & & \text{ATGGC} & & \text{ATGG-} \\
\end{align*}
\]

$M_0(i, j) = score(S[i], T[j]) + \max \begin{cases} 
M_0(i-1, j-1) \\
M_1(i-1, j-1) \\
M_2(i-1, j-1) 
\end{cases}$

$M_1(i, j) = \max \begin{cases} 
M_0(i, j-1) + a \\
M_1(i, j-1) + b \\
M_2(i, j-1) + a 
\end{cases}$

$M_2$ is similarly computed.
The goal of this paper is to give Hirschberg's idea the visibility it deserves by developing a linear-space version of Gotoh's algorithm.

where $\sigma_{\text{max}} = \max_{a,b} \sigma(a,b)$ (Smith et al., 1981). Thus, to produce an alignment that maximizes the similarity score, first apply these transformations and then run the program described in this paper with the resulting $w$, $g$ and $h$. If the minimum conversion score is $C$, then the corresponding maximum alignment score is $\frac{1}{2}(M + N)\sigma_{\text{max}} - C$.

Gotoh (1982) gave an algorithm that solves such problems in $O(MN)$ time. If only the minimum cost is desired, then it is easy to implement the algorithm in $O(N)$ space, where $N$ can be taken as the shorter sequence length. If one also desires a set of operations attaining the minimum cost, then straightforward implementations need $O(MN)$ space. In practice, this space

“The goal of this paper is to give Hirschberg’s idea the visibility it deserves by developing a linear-space version of Gotoh’s algorithm.”
Sequence Length

Number of nucleotide base pairs (or amino acid residues) in the sequence record. In this example, the sequence length is 5028 bp.

There is no maximum limit on the size of a sequence that can be submitted to GenBank. You can submit a whole genome if you have a contiguous piece of sequence from a single molecule type. However, there is a limit of 350 kb on an individual GenBank record (with some exceptions, as noted in section 1.3.2 of the release notes for GenBank 112.0). That limit was agreed upon by the international collaborating sequence databases to facilitate handling of sequence data by various software programs. (For more

Question: Why 350kb?
History Fact

5.25 inch DS/DD "Low Density" or Double Density

The older 5.25 inch "flexible floppy". Double sided, double density, usually formatted at 360K. These are only available FORMATTED at 360K. 5.25 floppy sleeves not included with floppies. Call if 5.25 sleeves are needed with your floppy disk.

Flat rate Priority Mail shipping in the US only $4.95.

Screenshot of floppydisk.com
Amino Acid Matrix

• Earlier justifications of the sequence alignment algorithms are more about protein sequences.
• Dayhoff developed the first amino acid substitution matrix. PAM.
• Cited 5257 times.
PAM: evolution distance of about 1 substitution per 100 amino acids.

Figure 78. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.

Figure 79. Matrix of accepted point mutations derived from the tree of Figure 78.
|   | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | Val |
| Ala | 9867 | 2 | 9 | 10 | 3 | 8 | -17 | 21 | 2 | 6 | 4 | 2 | 6 | 2 | 22 | 35 | 32 | 0 | 2 | 18 |
| Arg | 1 9913 | 3 | 0 | 10 | 0 | 0 | 10 | 3 | 1 | 19 | 4 | 1 | 4 | 6 | 1 | 6 | 1 | 8 | 0 | 1 |
| Asn | 41 9822 | 36 | 0 | 4 | 6 | 21 | 3 | 1 | 13 | 0 | 1 | 2 | 20 | 9 | 1 | 4 | 1 |
| Asp | 6 42 9859 | 0 | 0 | 6 | 53 | 6 | 4 | 1 | 0 | 3 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 1 |
| Cys | 1 1 0 0 0 | 1 0 | 0 | 973 | 0 | 0 | 1 | 2 | 0 | 1 | 4 | 1 | 0 | 3 | 4 | 2 | 0 | 1 | 2 |
| Gln | 3 9 4 5 | 0 9876 | 27 | 1 | 23 | 2 | 3 | 6 | 4 | 0 | 6 | 2 | 2 | 2 | 0 | 0 | 1 |
| Glu | 10 1 0 56 | 35 9866 | 4 | 2 | 3 | 1 | 4 | 1 | 0 | 3 | 4 | 2 | 0 | 1 | 2 |
| Gly | 21 1 1 12 | 11 | 1 | 3 | 7 9935 | 1 | 0 | 1 | 2 | 1 | 1 | 3 | 2 | 1 | 3 | 0 | 0 | 5 |
| His | 1 8 18 3 | 1 | 20 | 1 | 0 | 9912 | 0 | 1 | 1 | 0 | 2 | 3 | 1 | 1 | 1 | 1 | 4 | 1 |
| Ile | 2 2 3 1 | 2 | 1 | 2 | 0 | 0 | 9872 | 9 | 2 | 12 | 7 | 0 | 1 | 7 | 0 | 1 | 33 |
| Leu | 3 1 3 | 0 | 0 | 6 | 1 | 1 | 4 | 22 9947 | 2 | 45 | 13 | 3 | 1 | 1 | 3 | 4 | 2 | 15 |
| Lys | 2 37 28 6 | 0 | 12 | 2 | 7 | 2 | 2 | 4 | 1 | 9926 | 20 | 0 | 3 | 8 | 11 | 0 | 1 | 1 |
| Met | 1 1 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 5 | 8 | 4 9874 | 1 | 0 | 1 | 2 | 0 | 0 | 4 |
| Phe | 1 1 1 | 0 | 0 | 0 | 1 | 2 | 8 | 6 | 0 | 4 9946 | 0 | 2 | 1 | 3 | 2 | 8 | 0 |
| Pro | 13 5 2 | 1 | 1 | 8 | 3 | 2 | 5 | 1 | 2 | 1 | 1 | 9926 | 12 | 4 | 0 | 0 | 2|
| Ser | 28 11 34 | 7 | 11 | 4 | 6 | 16 | 2 | 2 | 1 | 7 | 4 | 3 | 17 | 9840 | 38 | 5 | 2 | 2 |
| Thr | 22 3 13 | 4 | 1 | 3 | 2 | 2 | 1 | 11 | 2 | 8 | 6 | 1 | 5 | 32 | 9971 | 0 | 2 | 9 |
| Trp | 0 2 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 9976 | 1 | 0 |
| Tyr | 1 0 3 | 0 | 3 | 0 | 1 | 0 | 4 | 1 | 1 | 0 | 0 | 2 | 1 | 0 | 1 | 1 | 2 | 9945 | 1 |
| Val | 13 2 1 | 1 | 3 | 2 | 2 | 3 | 3 | 57 | 1 | 1 | 17 | 1 | 3 | 2 | 10 | 0 | 2 | 9991 |

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, $M_{ij}$, gives the probability that the amino acid in column $j$ will be replaced by the amino acid in row $i$ after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.
### N PAM

| ORIGINAL AMINO ACID | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| Ala     | 13 | 6 | 9 | 3 | 5 | 8 | 9 | 12 | 6 | 8 | 6 | 7 | 7 | 4 | 11 | 11 | 11 | 12 | 2 | 4 | 9 |
| Arg     | 17 | 4 | 3 | 2 | 5 | 3 | 2 | 6 | 3 | 2 | 9 | 4 | 1 | 4 | 4 | 1 | 7 | 2 | 2 | 2 |
| Asn     | 4  | 4 | 6 | 7 | 2 | 5 | 6 | 4 | 6 | 3 | 2 | 5 | 3 | 2 | 4 | 5 | 4 | 2 | 3 | 3 |
| Asp     | 5  | 4 | 8 | 11| 1 | 7 | 10| 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| Cys     | 2  | 1 | 1 | 52| 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 4 | 2 | 3 | 3 |
| Gin     | 5  | 5 | 6 | 1 | 30| 7 | 3 | 7 | 2 | 3 | 5 | 3 | 1 | 4 | 3 | 3 | 1 | 2 | 3 |
| Glu     | 7  | 11| 1 | 9 | 12| 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| Gly     | 12 | 5 | 10| 10| 4 | 7 | 9 | 27| 5 | 5 | 4 | 6 | 5 | 3 | 8 | 11| 9 | 2 | 3 | 7 |
| His     | 5  | 5 | 4 | 2 | 7 | 4 | 2 | 15| 2 | 2 | 3 | 2 | 2 | 2 | 3 | 2 | 3 | 2 | 3 |
| Ile     | 3  | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 10| 5 | 2 | 6 | 5 | 2 | 3 | 4 | 1 | 3 | 9 |
| Leu     | 6  | 4 | 4 | 2 | 6 | 4 | 3 | 5 | 15| 34| 4 | 20| 13| 5 | 4 | 6 | 6 | 7 | 13|
| Lys     | 6  | 18| 10| 8 | 2 | 10| 8 | 5 | 8 | 5 | 4 | 24| 9 | 2 | 6 | 8 | 4 | 3 | 5 |
| Met     | 1  | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 6 | 2 | 1 | 1 | 1 | 1 |
| Phe     | 2  | 1 | 2 | 1 | 1 | 1 | 1 | 3 | 5 | 6 | 1 | 4 | 32| 1 | 2 | 2 | 4 | 20| 1 |
| Pro     | 7  | 5 | 5 | 4 | 3 | 5 | 4 | 5 | 3 | 3 | 4 | 3 | 2 | 20| 6 | 5 | 1 | 2 | 4|
| Ser     | 9  | 6 | 8 | 7 | 6 | 7 | 7 | 9 | 6 | 5 | 4 | 7 | 5 | 3 | 9 | 10| 9 | 4 | 4 |
| Thr     | 8  | 5 | 6 | 6 | 4 | 5 | 6 | 4 | 6 | 4 | 6 | 5 | 3 | 6 | 8 | 11| 2 | 3 | 6|
| Trp     | 2  | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 55| 1 | 0 |
| Tyr     | 1  | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 15| 1 | 7 | 2 | 3 | 31|
| Val     | 7  | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 15| 10| 4 | 10| 5 | 5 | 5 | 7 | 2 | 4 |

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala in the first sequence will contain Ala in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.

- The 1 PAM matrix can multiply itself $N$ times to get the $N$ PAM matrix.
BLOSUM


• That is when his wife, a mathematician by training, got involved. They started by constructing a database of conserved protein motifs, called BLOCKS. The Henikoffs then developed a technique to extract alignment “scores” from the BLOCKS database to determine how related two proteins might be. They called their technique BLOSUM, for BLOcks of Amino Acid SUbstitution Matrix.
They first estimated probabilities that two amino acids occur in the same column of a pair of alignment $P(x, y)$.

To avoid over-sampling of some proteins, they clustered the proteins too close (e.g. 62% or higher identity) into one group and count them with weight 1.

Then $\log \frac{P(x, y)}{P(x)P(y)}$ is the “log likelihood ratio” or sometimes “log odds” are calculated as the similarity score.
More Precisely

\[
score(x, y) = 2 \log_2 \frac{P_{xy}}{2P_x P_y}, \text{ if } x \neq y
\]

\[
score(x, x) = 2 \log_2 \frac{P_{xx}}{P_x P_x}
\]

In BLOSUM matrices these values are rounded to the nearest integer.
Practical Programs

• Computational biology was initially regarded as a branch of “theoretical computer science”.
• It provides 20 years of problems to work on.
• But from late 1980s to 1990, two practical programs started to change this...
Fastp

Rapid and sensitive protein similarity searches

DJ Lipman and WR Pearson

Abstract

An algorithm was developed which facilitates the search for similarities between newly determined amino acid sequences and sequences already available in databases. Because of the algorithm's efficiency on many microcomputers, sensitive protein database searches may now become a routine procedure for molecular biologists. The method efficiently identifies regions of similar sequence and then scores the aligned identical and differing residues in those regions by means of an amino acid replacability matrix. This matrix increases sensitivity by giving high scores to those amino acid replacements which occur frequently in evolution. The algorithm has been implemented in a computer program designed to search protein databases very rapidly. For example, comparison of a 200-amino-acid sequence to the 500,000 residues in the National Biomedical Research Foundation library would take less than 2 minutes on a minicomputer, and less than 10 minutes on a microcomputer (IBM PC).
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1984</td>
<td>On January 24, 1984 the Apple Macintosh is introduced.</td>
</tr>
<tr>
<td>1984</td>
<td>Steve Jobs and Steve Wozniak are awarded the National Medal of Technology</td>
</tr>
<tr>
<td>1984</td>
<td>Bill Gates is featured on the cover of TIME magazine.</td>
</tr>
<tr>
<td>1984</td>
<td>ASN.1 is first defined.</td>
</tr>
<tr>
<td>1984</td>
<td>The 3.5-inch floppy diskette is introduced and later becomes an industry standard.</td>
</tr>
<tr>
<td>1984</td>
<td>Dell Computer is founded May 3, 1984 in Austin Texas.</td>
</tr>
<tr>
<td>1985</td>
<td>Microsoft and IBM begin collaboration on the next-generation operating system (OS/2).</td>
</tr>
<tr>
<td>1985</td>
<td>Microsoft Windows 1.0 is introduced in November, 1985 and is initially sold for $100.00.</td>
</tr>
<tr>
<td>1985</td>
<td>ATI is founded.</td>
</tr>
</tbody>
</table>
FASTA


• “The FASTA program can search the NBRF protein sequence library (2.5 million residues) in less than 20 min on an IBM-PC microcomputer”.

• FASTA is also known as a sequence file format.
Basic Local Alignment Search Tool

Stephen F. Altschul¹, Warren Gish¹, Webb Miller²
Eugene W. Myers³ and David J. Lipman¹

¹National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

²Department of Computer Science
The Pennsylvania State University, University Park, PA 16802, U.S.A.

³Department of Computer Science
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990: accepted 15 May 1990)

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST),
directly approximates alignments that optimize a measure of local similarity, the maximal
segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP
scores allow an analysis of the performance of this method as well as the statistical
significance of alignments it generates. The basic algorithm is simple and robust; it can be
implemented in a number of ways and applied in a variety of contexts including straightforward
DNA and protein sequence database searches, motif searches, gene identification
searches, and in the analysis of multiple regions of similarity in long DNA sequences. In
addition to its flexibility and tractability to mathematical analysis, BLAST is an order of
magnitude faster than existing sequence comparison tools of comparable sensitivity.

• Cited by 40501
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>In 1990 Tim Berners-Lee, working with Robert Cailliau at CERN propose a 'hypertext' system, which is the first start of the Internet as we know it today.</td>
</tr>
<tr>
<td>1990</td>
<td>Microsoft releases Windows 3.0 a completely new version of Microsoft Windows. The version will sell more than 3 million copies in one year.</td>
</tr>
<tr>
<td>1990</td>
<td>Microsoft exceeds $1 billion in sales and becomes the first company to do so.</td>
</tr>
<tr>
<td>1990</td>
<td>Godwin's Law is conceived.</td>
</tr>
<tr>
<td>1990</td>
<td>Electronic Frontier Foundation or EFF is founded February 16, 1990.</td>
</tr>
<tr>
<td>1990</td>
<td>Hubble telescope goes into space.</td>
</tr>
<tr>
<td>1990</td>
<td>Microsoft releases its first product for the Russian market Russian DOS 4.01.</td>
</tr>
<tr>
<td>1990</td>
<td>The World, the first commercial Internet dial-up access provider comes online.</td>
</tr>
<tr>
<td>1990</td>
<td>Norton sells his software business to Symantec.</td>
</tr>
<tr>
<td>1990</td>
<td>Creative Labs Introduces the SoundBlaster Pro.</td>
</tr>
<tr>
<td>1990</td>
<td>Quarterdeck releases its memory management program QEMM386 version 5.1 which quickly becomes the fastest-selling software program in the United States.</td>
</tr>
<tr>
<td>1990</td>
<td>The Multimedia Personal Computer (MPC) standards are developed by Tandy and Microsoft.</td>
</tr>
<tr>
<td>1990</td>
<td>Microsoft and IBM stop working together to develop operating systems.</td>
</tr>
<tr>
<td>1990</td>
<td>IBM introduces XGA.</td>
</tr>
<tr>
<td>1990</td>
<td>ARPANET replaced by NSFNET.</td>
</tr>
<tr>
<td>1990</td>
<td>The first search engine Archie, written by Alan Emstage, Bill Heelan, and Mike Parker at McGill University in Montreal Canada is released on September 10, 1990.</td>
</tr>
<tr>
<td>1990</td>
<td>GSM standard is defined.</td>
</tr>
<tr>
<td>1990</td>
<td>The NIMH battery begins being used for commercial use.</td>
</tr>
<tr>
<td>1990</td>
<td>Panda Software is founded.</td>
</tr>
<tr>
<td>1990</td>
<td>Archie, the first search engine is introduced on September 10, 1990.</td>
</tr>
<tr>
<td>1990</td>
<td>Gopher is developed at the University of Minnesota. The program is a menu-driven search-and-retrieval tool and helps Internet users location information online.</td>
</tr>
<tr>
<td>1990</td>
<td>The Internet Movie Database (IMDb) is launched October 17, 1990.</td>
</tr>
<tr>
<td>1990</td>
<td>Intel releases the 80386SL processor that uses low power and found in many portable computers.</td>
</tr>
<tr>
<td>1990</td>
<td>Tim Berners-Lee successfully sets up the first web server at info.cern.ch on December 25, 1990.</td>
</tr>
</tbody>
</table>
BLAST finds “hits” then extends

```
GCNTACACGTCACCACATGTTGCTGACCACACCACNCATGTCTCTAGTGATCCCTCATAAGTTCCCAACAAAGTTTGC
```

```
GCCTACACACCAGCTTTGAT-GTCTCTATGTGATCCCTGAAAAGTTCCAGCGTATTTTGC
```

![Diagram showing sequence alignment](image)

- Time complexity becomes $O(4^{-11}mnT)$
- The homology grows from the 11 consecutive matches therefore they are called the **seed**
- More precisely these are called hits. A seed often is used to refer to the way these 11 positions are selected.
For Protein Sequence

• A hit is the approximate match of two 3-mers, so that the total score (under PAM) is above a threshold.

• Use query sequence’s hitting 3-mers to build a DFA (Deterministic Finite Automata). Use the DFA to scan the database.
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

Stephen F. Altschul*, Thomas L. Madden, Alejandro A. Schäffer¹, Jinhui Zhang, Zheng Zhang², Webb Miller² and David J. Lipman

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, ¹Laboratory of Genetic Disease Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA and ²Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA

Received June 20, 1997; Revised and Accepted July 16, 1997

• Cited by 41240
New Ideas in Second BLAST Paper

• Two-hits:

• Position-Specific Iterated (PSI):
  – Multiple sequence alignment of returned proteins.
  – Calculate position-specific scoring matrix (instead of using BLOSUM).
  – Search again with the new matrix.
Computer History for 1997

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>IEEE releases 802.11 (WiFi) standard.</td>
</tr>
<tr>
<td>1997</td>
<td>Yahoo! introduces Yahoo Mail.</td>
</tr>
<tr>
<td>1997</td>
<td>Intel Pentium II is introduced on May 7, 1997.</td>
</tr>
<tr>
<td>1997</td>
<td>IBM's Deep Blue computer defeats world champion chess player Garry Kasparov May 11, 1997 in their second six-game showdown, winning the tie-breaking game in only 62 minutes.</td>
</tr>
<tr>
<td>1997</td>
<td>Perl 5.004 is released May 15, 1997.</td>
</tr>
<tr>
<td>1997</td>
<td>Kaspersky is founded.</td>
</tr>
<tr>
<td>1997</td>
<td>Bill Gates is now the world's richest businessman.</td>
</tr>
<tr>
<td>1997</td>
<td>Microsoft saves Apple with a $150 million investment August 6, 1997.</td>
</tr>
<tr>
<td>1997</td>
<td>The google.com domain name is registered after Sergey Brin and Larry Page decide to change the name of their BackRub search engine to Google September 15, 1997.</td>
</tr>
<tr>
<td>1997</td>
<td>Microsoft acquires Hotmail a free e-mail service in December 1997.</td>
</tr>
</tbody>
</table>
Example of missing a target

• Fail:
  – GAGTACTCAACACACAACTATAGTGGGCAATGGAAAAT
  – || || || || || || || || || || || || || || || || ||
  – GAATACTCAACAGCAACATCAATGGGCAGCAGAAAAT

• Two objectives
  – Sensitivity (success rate of finding a homology)
    • needs a shorter seed
  – Speed (the fewer hits, the faster)
    • needs a longer seed

• But we really want both!
PatternHunter: faster and more sensitive homology search

Bin Ma\textsuperscript{1}, John Tromp\textsuperscript{2} and Ming Li\textsuperscript{3}

\textsuperscript{1}Computer Science Department, University of Western Ontario, London N6A 5B8, Canada, \textsuperscript{2}Bioinformatics Solutions Inc., 145 Columbia Street West, Waterloo, Ont. N2L 3L2, Canada and \textsuperscript{3}Computer Science Department, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada and Bioinformatics Lab, Computer Science Department, University of California, Santa Barbara, CA 93106, USA

Received on August 24, 2001; revised on October 10, 2001; accepted on October 15, 2001
Small change

• We now use eleven “spaced” matches as a hit
  – compare to consecutive matches
• Match locations given by a pre-defined pattern
  – 111010010100110111
  – Eleven required positions (weight=11)
  – Seven “don’t care” positions
Difference

- **Speed**
  - Random hit probability is still $4^{-11}$
  - “Weight” of the seed determines the hit probability

- **Sensitivity**
  - Spaced seed longer, so fewer positions to hit
  - Intuitively lower.

```
1 111111111111110100110111
1 111010010100110111
```
Simulated sensitivity curves
Why spaced seeds are better?

- BLAST’s seed usually uses more than one hits to detect one homology (redundant)
- Spaced seeds uses fewer hits to detect one homology (efficient)
Seed shape is important

• PH’s seed not overlap much when slides:

```
111010010100110111
111010010100110111
111010010100110111
```

• The hits at different positions are independent.
2004

Multiple Spaced Seeds

• Different spaced seeds detect different homologies.

• Use k seeds simultaneously!
  – Slows down k times.

• Is this worth?
  – Decreasing the seed weight can also increase the sensitivity.
  – Slows down 4 times.
Simulated sensitivity curves:

- Solid curves: Multiple (1, 2, 4, 8, 16) weight-12 spaced seeds.
- Dashed curves: Optimal spaced seeds with weight = 11, 10, 9, 8.
- Typically, “Doubling the seed number” gains better sensitivity than “decreasing the weight by 1”.