Implementing bioinformatics algorithms in TEX

Gotoh package, a case study

Takuto ASAKURA (wtsnjp)

The University of Tokyo

TUG@BachoTEX 2017

Biological sequences

DNA, RNA, Amino acids, etc.

Biologists want to know the degree of *similarity* among 2 or more sequences.

Pairwise sequence alignment

The problem

Input: Two biological sequences

$$A \equiv a_1 a_2 a_3 \dots a_m, \qquad B \equiv b_1 b_2 b_3 \dots b_n$$

where a_i and b_j are chosen from a finite alphabet, e.g. $\{\mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}\}$.

Output: An alignment between A and B.

Examples

```
bachotex   context    bioinformatics
|||||||*      |||     ||     **|| **
bachotek   ---tex-    bi-----blat-ex
```

| *match* $\ast$ *mismatch*   - *gap*

Longest Common Subsequence (LCS)

LCS problem

- Want to get the LCS of *A* and *B*
- A simplest form of sequence alignment
- Score 1 for *matches* and 0 for *gaps*

The solution

$$s_{i,j} = \max \begin{cases} s_{i-1,j} \\ s_{i,j-1} \\ s_{i-1,j-1} + 1 \end{cases}$$

The Gotoh algorithm: DP

Sequence alignment has a slightly more complex scoring scheme.

Example

$\mathrm{match} = 1,\ \mathrm{mismatch} = -1,\ g(l) = -d - (l-1)e$

The algorithm

Sequence alignment in $O(mn)$ time:

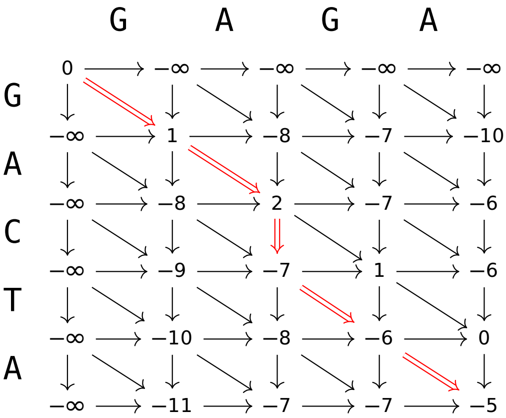
$$M_{i+1,j+1} = \max\left\{M_{ij}, I_{x\,ij}, I_{y\,ij}\right\} + c_{a_i b_j}$$

where

$$I_{x\,i+1,j} = \max\left\{M_{ij} - d, I_{x\,ij} - e, I_{y\,ij} - d\right\},$$
$$I_{y\,i,j+1} = \max\left\{M_{ij} - d, I_{y\,ij} - e\right\}.$$

The Gotoh algorithm: trace back

Start at maximum entry, trace back to first entry.



		G	A	G	A
	0	−∞	−∞	−∞	−∞
G	−∞	1	−8	−7	−10
A	−∞	−8	2	−7	−6
C	−∞	−9	−7	1	−6
T	−∞	−10	−8	−6	0
A	−∞	−11	−7	−7	−5

```
GACTA
GA-GA
```

What I did is . . .

LaTeX package

TeX A Turing Machine

LaTeX Widely used for typesetting papers

+

The Gotoh algorithm

- Can be written in short code
- Calculated with limited range of numbers
- Produces visual results

The **Gotoh** package

Usage

- `\Gotoh`{⟨*sequence A*⟩}{⟨*sequence B*⟩}
  - Executes the algorithm
  - Returns the results to specified CSs
- `\GotohConfig`{⟨*key-value list*⟩}
  - Setting various parameters
  - e.g. algorithm parameters, CSs to store results

Example

Input:

```
\Gotoh{ATCGGCGCACGGGGGA}
      {TTCCGCCCACA}
\texttt{\GotohResultA} \\
\texttt{\GotohResultB}
```

Output:

```
ATCGGCGCACGGGGGA
TTCCGCCCAC.....A
```

Combining with TEXshade

The TEXshade package

- A part of BioTEX, produced by Eric Beitz
- Shading and labeling *preprocessed* alignments
- Can be used to format the outputs of **Gotoh**

Example

```
\newcommand{\PrintAlignment}[3][\relax]{%
  \Gotoh{#2}{#3}%
  \immediate\openout\FASTAfile=\jobname.fasta
  \writeFASTA{> Seq 1^^J\GotohResultA}%
  \writeFASTA{> Seq 2^^J\GotohResultB}%
  \immediate\closeout\FASTAfile
  \texshade{\jobname.fasta}#1\endtexshade}
```

Let me show you a demonstration!

Features and future

Advantages

The **Gotoh** package is:

- simple to use
- long-lasting
- cross-platform

Future work

- Preparing the documentation
- Uploading to CTAN
- Adding some functions such as:
  - showing edit graphs
  - calculating multiple alignment (≥ 3 sequences)

Conclusion

Algorithms in any field which are:

- often used for creating documents
- easy to implement

are worth implementing in TEX.

Example

`diff` function for **listings**

Thank you & Happy TEXing!!