Sequence analysis

CTX-BLAST: context sensitive version of protein BLAST

Anna Gambin* and Piotr Wojtalewicz
Institute of Informatics, Warsaw University, Banacha 2, 02-097, Warsaw, Poland

Received on November 27, 2007; revised on March 13, 2007; accepted on April 3, 2007
Advance Access publication April 25, 2007
Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: We present a software tool CTX-BLAST that incorporates contextual alignment model into the popular protein BLAST program. Our alignment tool allows us to investigate the effect of context-dependency in the protein alignment much more efficient than using previous dynamic algorithms. The software makes use of non-symmetric contextual substitution tables and calculates the statistical significance of a given alignment according to the contextual statistical model.

Availability: CTX-BLAST is an open source software freely available from www.sourceforge.net/projects/CTX-BLAST. A program for statistical estimation of E-value parameters and the contextual substitution table CTX-BLOSUM62 are also provided.

Contact: aniag@mimuw.edu.pl

Supplementary information: Supplementary data are available at Bioinformatics online.

1 CONTEXTUAL ALIGNMENT

The model of contextual alignment of biological sequences has been introduced in Gambin et al. (2002). It is an extension of the classical alignment, in which the cost of an amino acid substitution depends on the surrounding residues. Consequently, in this model the cost of transforming one protein sequence into another depends on the order of editing operations. In the following example we see that the relative order of two substitutions (H → P and V → D) applied to the same sequence affects the score, if a contextual scoring function is used (substitution scores are taken from CTX-BLOSUM62).

There exist dynamic algorithms for calculating the maximal global and local contextual alignment scores (Gambin et al., 2002). Their complexity is (up to a constant factor) the same as in the classical non-contextual alignment model (i.e. Needleman–Wunsch and Smith–Waterman algorithms), if the insertion and deletion score functions are affine.

The exhaustive contextual dynamic approach is too slow for searching large genomic database. On the other hand, popular BLAST tool (Altschul et al., 1990) emphasizes speed over sensitivity and is very efficient. Our ultimate goal was to increase the sensitivity of BLAST by incorporating the contextual model and keeping its efficiency at the same time.

Our extension of BLAST, called CTX-BLAST, allows us to investigate context sensitivity by aligning a query sequence against large databases. Using our new tool, for the first time we can detect subtle context-sensitive homologies, which were missed by the standard BLAST algorithm.

2 CTX-BLAST ALGORITHM

We have modified the open source code of NCBI BLAST distribution (www.ncbi.nlm.nih.gov/BLAST). Most important changes include the gap extension function and the alignment reconstruction. Both tasks in CTX-BLAST are performed in the contextual manners i.e. using the dynamic programming algorithms from Gambin et al. (2002) and the contextual substitution table. The construction of context dependent substitution tables is discussed in Gambin et al. (2006).

The crucial point of the BLAST algorithm is the calculation of statistical significance of the alignment. According to Gambin et al. (2006), statistics of optimal non-gapped contextual alignment follows the same extreme value distribution as in the non-contextual case (Altschul and Gish, 1990). Hence, we have decided to adopt the island method (Altschul et al., 2001) for estimation of the parameters $K$ and $\lambda$ required for E-value calculation for the alignment score $S$ of two sequences of length $m$ and $n$, respectively:

$$E = \text{value}(S) = Kmme^{-\lambda S}.$$  

The goal is to choose the threshold $c$ (i.e. the score of the island) that results in the best estimation of $\lambda$ and $K$. When $c$ increases, the value of $\lambda_c$ decreases until the standard error masks all further changes of $\lambda_c$. As suggested in Altschul et al., (2001),
we have chosen the threshold $c = 44$ determined in this way and obtained $\lambda = 0.2110$ and $K = 0.008$ (standard error <1%). Our program also estimates parameters $\alpha$ and $\beta$ necessary for the correction of the edge effect.

Table. Estimated values for $K$ and $\lambda$ parameters

| $c$ | $R_c$ | $\lambda_c$ | $K_c$ |
|-----|-------|-------------|-------|
| 40  |        | 0.2122±0.0010 (0.49%) | 0.009 |
| 41  |        | 0.2120±0.0012 (0.55%) | 0.009 |
| 42  |        | 0.2110±0.0013 (0.61%) | 0.008 |
| 43  |        | 0.2104±0.0014 (0.68%) | 0.008 |
| 44  | **19** | **0.2110±0.0016 (0.76%)** | **0.008** |
| 45  |        | 0.2105±0.0018 (0.84%) | 0.008 |
| 46  |        | 0.2092±0.0020 (0.94%) | 0.008 |
| 47  |        | 0.2079±0.0022 (1.04%) | 0.007 |
| 48  |        | 0.2084±0.0024 (1.15%) | 0.007 |
| 49  |        | 0.2000±0.0027 (1.28%) | 0.008 |
| 50  |        | 0.2118±0.0030 (1.41%) | 0.009 |

### 3 EXPERIMENTS

All experiments described here were performed with the use of contextual version of BLOSUM62 substitution table called CTX-BLOSUM62 and affine gap penalty $-\alpha + \beta k$, where $k$ is the length of the gap.

The ultimate objective of our research was to verify whether the use of CTX-BLAST may help in homology detection. We compared the performance of our tool with standard BLAST using carefully selected dataset (Kann, 2003). It consists of 100 pairs of homologous proteins from PFAM database (Finn et al., 2006), and 1000 pairs from COG database (Tatusov et al., 1997). All proteins are remote homologs with low sequence similarity and well-established structural relationships. For PFAM dataset the CTX-BLAST outperformed BLAST in 66% of cases, by yielding longer alignment, or by detecting significant similarity while BLAST failed. In the case of COG dataset, the CTX-BLAST succeeded in 40% of cases (note, however, that the structural similarity is less evident than for the PFAM dataset). Below we present in detail an example of Glycoprotein PFAM family. The complete output of the experiment can be found at the Supplementary web page: www.mimuw.edu.pl/~aniag/CTX-BLAST.

Two sequences were taken from Spike glycoprotein family PF09794. Using these sequences as queries to PDB database (Berman et al., 2000), we have determined the best structural hits. Both sequences exhibit significant structural similarity to the G protein chain. The analyzed PDB alignments with G-protein cover almost whole query sequences.

The glycoprotein spike is made up of a trimer of G proteins. The figure 2 presents the crystal structure of G protein spike determined recently (Roche et al., 2006). Three G proteins are visible. BLAST alignment for our pair of sequences is only 198 residues long:

```
Score = 38.1 bits (87)
Id. = 43/198 (21%), Pos. = 83/198 (41%), Gap = 16/198 (8%)
```

The CTX-BLAST resulted in longer (353 residues) and more significant alignment, that covers the whole structural motif.

```
Score = 48.8 bits (82)
Id. = 56/353 (15%), Pos. = 111/353 (31%), Gap = 25/353 (7%)
```

Another experiment performed aimed in detecting the group of context sensitive proteins. We used two artificial amino acid sequences that give much higher score while contextually aligned (we call them bait sequences). The real proteins that showed high similarity to bait sequences were found. These proteins were used further to scan the COG database. The results can be found at Supplementary web page. The most interesting outcome of this experiment is the existence of large group of proteins, that exhibit unusual context sensitivity. The underlying molecular phenomena is currently under study.

### 4 USAGE

The following command runs the standard version of BLAST:

```
blast -p program -d database -i query \
-o output_file
```

where program is in our case blastp (i.e. protein version of BLAST tool), database is the name of database to scan for significant alignments and query is the query sequence.

The contextual version of BLAST requires two additional parameters: $C$ defines the contextual substitution table CTX-BLOSUM62, the contextual counterpart of BLOSUM62 and $x$ defines the file containing the parameters for statistical significance calculation in the following order: $\alpha$, $\beta$, $\lambda$ and $K$. 

Figure 2. The crystal structure of G protein spike from Roche et al., 2006.
The exemplary usage of CTX-BLAST on the COG database kyva is the following:

```
blast -p blastp -d kyva -i query.txt
-o outputctxb.htm -C CTX-BLOSUM62 -x stat.txt
```

To estimate the parameters for statistical significance we have to run program island, which implements the island method in the contextual model (the program is freely available together with CTX-BLAST). Here is the example invocation:

```
island --phi=phi.txt --matrix=CTX-BLOSUM62
--align=LOCAL --wordsize=7000
--islecutoff=20 --iterisle=250
--gapopen=11 --gapcont=1 --output=scores.txt
--border=1000
```

The parameter `wordsize` defines the length of compared sequences and `border` is the frame width, see Altschul et al. (2001) for details. The gap penalty is defined by parameters `gapopen` and `gapcont`. The number of comparisons to perform is `iterisle`, and `islecutoff` is the threshold value for the islands. The output parameter specifies the output file. From statistics stored in this file (e.g. `scores.txt`), we estimate parameters $\lambda$, $K$, $\alpha$ and $\beta$ using the R-script `isle.R` (freely available at www.sourceforge.net/projects/CTX-BLAST).

**ACKNOWLEDGEMENTS**

The research described in this paper was partially supported by Ministry of Science and Higher Education grant 3 T11F 021 28.

*Conflict of Interest*: none declared.

**REFERENCES**

Altschul, S.F. and Gish, W. (1996) Local alignment statistics. *Meth. Enzymol.*, 266, 460–480.

Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–410.

Altschul, S.F. *et al.* (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, 29, 351–361.

Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, 28, 235–242.

Finn, R.D. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, 34, D247–D251.

Gambin, A. *et al.* (2002) Contextual Alignment of Biological Sequences. *Bioinformatics*, 18, 116–127.

Gambin, A. *et al.* (2006) Context dependent alignment: a new method for comparing biological sequences. *J. Comput. Biol.*, 13, 81–101.

Kann, M. (2003) Private communication.

Roche, S. *et al.* (2006) Crystal structure of the low-ph form of vesicular stomatitis virus glycoprotein G. *Science*, 313, 187–191.

Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, 278, 631–637.