BIOINFORMATICS APPLICATIONS NOTE
Vol. 20 no. 15 2004, pages 2488–2490
doi:10.1093/bioinformatics/bth265

Sfixem—graphical sequence feature display in Java

Alistair M. Chalk*, Martin Wennerberg and Erik L. L. Sonnhammer

Center for Genomics and Bioinformatics, Karolinska Institutet, 17177 Stockholm, Sweden

Received on October 16, 2003; revised and accepted on March 31, 2004
Advance Access publication April 15, 2004

ABSTRACT
Summary: Sfixem is an sequence feature series (SFS) visualization tool implemented in Java. It is designed to visualize data from sequence analysis programs, allowing the user to view multiple sets of computationally generated analysis to assist the analysis process. SFS is used as the data exchange format.

Availability: Sfixem is available for direct usage or download for local usage at http://sfixem.cgb.ki.se. A protein sequence analysis workbench using Sfixem is available at http://sfinx.cgb.ki.se.

Contact: Alistair.Chalk@cgb.ki.se

INTRODUCTION

Sequence analysis can be greatly improved by using graphical tools. Such tools allow expert human interpretation of algorithmic analysis, often from multiple sources. Having all the relevant data and analysis output available in a graphical form simplifies the analysis process, increasing the reliability of human interpretation.

Several systems exist that can be used for sequence segment display. Most of these are integrated systems that are fairly large, often involving substantial frameworks for complex data types. These systems include ChromoScope (Zhang *et al*., 1994), bioWigets (Searls, 1995), APIC (Bisson and Garreau, 1995), the BDGP java sequence viewer (Rubin, 1996), GAIA (Bailey *et al*., 1998), ACEDB (Eeckman and Durbin, 1999) and Artemis (Rutherford *et al*., 2000). We consider the investment of time required to generate and parse complex data types (e.g. ASN.1) excessive, and for the data we display (segments and curves), a simple tabular format is sufficient. We therefore use the sequence feature series (SFS) format (Sonnhammer and Wootton, 2001), which we briefly describe below. To make Sfixem more useful, we plan to make it compatible with data standards, such as DAS and XML in the future.

Two existing viewers that handle curves and segment data using SFS are Blixem (Sonnhammer and Durbin, 1994) and Dotter (Sonnhammer and Durbin, 1995). Both tools were primarily designed to display specialized sequence analysis data, with added SFS viewing functionality. A disadvantage with these and many other tools is the need to install the software locally. Local software installation, while trivial, is often a factor in tools being under-utilized by the biological community. Reasons for this can be lack of administrator rights on the machine, operating system incompatibility or simply reluctance to install external software.

SFIXEM

The visualization tool Sfixem was designed to visualize curve and segment data relating to a single sequence, considering SFS format data as input from any sequence analysis tool. In contrast to Blixem and Dotter, where SFS display capabilities were added to existing specialized analysis tools, Sfixem was written from scratch with generalized SFS display as the single purpose. It is hence a cleaner implementation and more ergonomic to use. The implementation as a Java applet removes the need to install the program locally, the applet requires the Java 2 Runtime Environment version 1.4.

The viewer has zoom functionality with variable amount of detail shown (i.e. coloured boxes or individual residues) dependent on zoom level. The user can get an overall view and zoom in on areas of interest to examine details at the single residue level. For segments, the residues will be shown in the coloured boxes when zoomed into sufficient detail. Additional functionality includes a series picker—allowing the user to select only a subset of the full amount of analysis data to view, and the modification of curve and the segment height.

DATA FORMAT

The SFS data format contains two primary data types that are used by most prediction programs: segments and curves. The segment type is similar to the popular GFF format (http://www.sanger.ac.uk/Software/formats/GFF/); GFF format can in fact be embedded in the SFS data. However, SFS also allows other types of data, for instance for drawing curves which is

---

*To whom correspondence should be addressed.

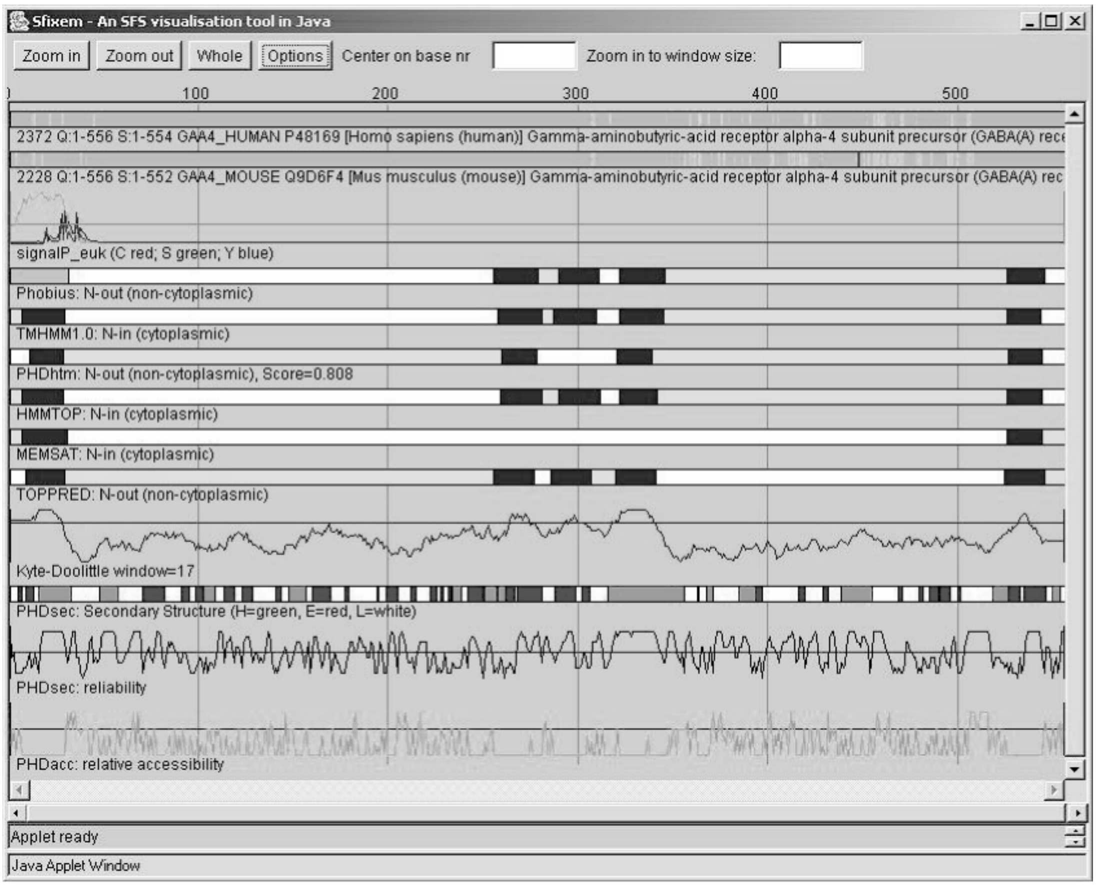

Fig. 1. Sfixem analysis of GAA4_BOVIN. The two rows at the top show homologies, where matching residues are shown in bright blue, conserved substitutions are in dark blue and mismatches are in grey. This is followed by signal peptide and transmembrane predictions, where TM segments are shown in brown.

often more illustrative rather than segments. There is also a native support for passing the sequence itself in the SFS file, which is useful for many visualization tasks. The simplicity of the SFS format minimizes the overhead for both the analysis software and the visualization tool to achieve graphical and user-friendly display of results. A full specification of the SFS format is given in Sonnhammer and Wootton (2001) and on the Sfixem website (see Availability). We here include a brief description on some definitions in SFS to illustrate the basic features. The SFS data are in Courier font. Data fields enclosed by ‘<>’ must be single words, while data fields enclosed by ‘[]’ may consist of multiple words.

Header

```
# SFS format 1.0
```

Sequence data

```
# SFS type=SEQ <sequence> <seqname>
  [annotation]
```

Segment data

```
# SFS type=SEG
<score><seqname><seriesname>
  <start><end> <look> [annotation]
```

XY data (curves)

```
# SFS type=XY <seqname> <seriesname>
  <look> [annotation]
<x-value> <y-value>
<x-value> <y-value>
```

HSP data

```
# SFS type=HSP
<score><qname><qframe><qstart><qend>
  <sname><sframe><sstart><send>
  <sequence>
```

An example of Sfixem usage is shown in Figure 1, in which a collection of transmembrane (TM) topology, signal peptide, and structure predictions for GAA4_BOVIN (Swissprot:P20237) are viewed, along with homology matches (produced by the Sfinx server, http://sfinx.cgb.ki.se). Visualizing collections of predictions from different algorithms allows the user to make a more informed decision based on all available data. For instance, it is quite clear that the PhD and Memsat predictions are missing internal TM segments in this case, and that the N-terminal TM segment predicted by most methods is really a signal peptide.

ACKNOWLEDGEMENT

This work was made possible by a grant from Pfizer Inc.

REFERENCES

Bailey,L.C.,Jr, Fischer,S., Schug,J., Crabtree,J., Gibson,M. and Overton,G.C. (1998) GAIA: framework annotation of genomic sequence. *Genome Res.*, **8**, 234–250.

Bisson,G. and Garreau,A. (1995) APIC: a generic interface for sequencing projects. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 57–65.

Eeckman,F.H. and Durbin,R. (1995) ACeDB and macace. *Methods Cell Biol.*, **48**, 583–605.

Rubin,G. (1996) Around the genomes: the Drosophilia Genome Project. *Genome Res.*, **6**, 71–79.

Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M., and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

Searls,D.B. (1995) bioTk:componentry for genome informatics graphical user interfaces. *Gene*, **163**, GC1–GC16.

Sonnhammer,E.L.L. and Durbin,R. (1994) A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.*, **10**, 301–307.

Sonnhammer,E.L.L. and Durbin,R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–GC10.

Sonnhammer,E.L. and Wootton,J.C. (2001) Integrated graphical analysis of protein sequence features predicted from sequence composition. *Proteins*, **45**, 262–273.

Zhang,J., Ostell,J. and Rudd,K. (1994) ChromoScope: a graphical interactive browser for *E. coli* data expressed in the NCBI data model. *Proceedings of the 27th Annual Hawaii International Conference on System Sciences*, (HICSS 1994), Maui, Hawaii, 4–7 January. IEEE Computer Society, Vol. 5, pp. 58–67.