3 Biotech (2015) 5:87–92
DOI 10.1007/s13205-014-0222-0

ORIGINAL ARTICLE

An integrated Java tool for generating amino acid sequence alignments with mapped secondary structure elements

Conan K. Wang · Andreas Hofmann

Received: 2 April 2014 / Accepted: 29 April 2014 / Published online: 20 May 2014

Abstract The mapping of secondary structure elements onto amino acid sequences enhances the quality of alignments frequently used in phylogenetic, genomic and transcriptomic studies, as well as in molecular modelling. Here, we report recent updates to the Java application SBAL, an integrated tool to generate, edit, visualise and analyse secondary structure-based sequence alignments. The main goal of the software is to streamline the work flow in generation of structure-based alignments, and we have thus implemented the option to import and visualise sequence and structure information directly from any PDB file. The new feature is achieved by a Java application named ASSP which follows the original framework of the well-established dictionary of protein secondary structure by Kabsch and Sander. ASSP is also available as a stand-alone application. Other major additions to SBAL include the calculation of distance matrices, peptide properties, as well as detailed on-line tutorials for typical applications.

Keywords Modelling · Structure analysis · Structure-based amino acid sequence alignments

Introduction

The rate of completion of new genome sequencing projects has been rapidly increasing in the recent past, thus providing large amounts of information on new proteins. To characterise and classify the immense number of new proteins from genome and transcriptome projects, automated assignment methods are used that, in the majority of cases, correctly annotate a new protein sequence to a known homologous protein fold (Cantacessi et al. 2010, 2014). Assembled nucleotide sequences from genomic and transcriptomic studies are usually conceptually translated into predicted proteins using algorithms that identify protein-coding regions. The predicted peptide sequences are then analysed for protein identity, for example with the software InterProScan (Hunter et al. 2012), by comparison of sequences with data available in public databases, to infer known protein domains. However, this first-pass annotation does not reveal any specific molecular features of the annotated proteins, such as conservation of active sites, variations of the conserved fold (Osman et al. 2012), or novel structural elements (Cantacessi et al. 2013).

To gain more detailed insights at the molecular level, in the absence of an experimental three-dimensional structure, comparative modelling can be employed. In many cases, this results in generation of three-dimensional atomic models of the target protein. Frequently, however, essential details can be gleaned from appropriate amino acid sequence alignments. In this context, informed sequence alignments are essential for constructing motifs, profiles and atomic modelling instructions (Eidhammer et al. 2000;

Electronic supplementary material The online version of this article (doi:10.1007/s13205-014-0222-0) contains supplementary material, which is available to authorized users.

C. K. Wang
Institute for Molecular Biosciences, University of Queensland,
St Lucia, QLD, Australia

A. Hofmann
Structural Chemistry Program, Eskitis Institute, Griffith
University, Brisbane, QLD, Australia

A. Hofmann (✉)
Faculty of Veterinary Science, The University of Melbourne,
Parkville, VIC, Australia
e-mail: a.hofmann@griffith.edu.au

Hubbard and Blundell 1987; Marchler-Bauer et al. 2002; Sauder et al. 2000). Often, amino acid sequence identities between two distantly related proteins are rather low when comparing, for example, parasite with vertebrate proteins (A Jex & RB Gasser, pers. commun.), although there are a few exceptions (Hewitson et al. 2009). In such situations, it is difficult to obtain meaningful alignments based on amino acid sequence similarity alone (Marchler-Bauer et al. 2002; Sauder et al. 2000). Since the main criterion for structural homology of two proteins is that they adopt the same fold, structure-based amino acid sequence alignments have been used as the gold standard for sequence alignment evaluation (Hubbard and Blundell 1987; Russell and Barton 1994).

We have previously developed the Java application SBAL (Wang et al. 2012) to fill an apparent gap in the seamless transition from secondary structure-based sequence alignments to the visualisation of the results. The main emphasis in the development of this software was the ease-of-use and the integration of transitional steps such as reformatting and editing, integrated with aids for visualisation and analysis. Among the variety of input formats for the original SBAL software, sequence and secondary structure information could be read directly from experimental three-dimensional structures in protein data bank (PDB) format. This could be achieved either by reading information from the PDB header section (‘HELIX’ and ‘SHEET’ records) or externally processing the PDB file with the established software DSSP (Kabsch and Sander 1983), which assigns secondary structure information to each residue based on hydrogen bonding patterns. In many practical environments, however, we found that not all structures in PDB format contain ‘HELIX’/‘SHEET’ records (e.g. if they are not obtained from the PDB, but become accessible through ongoing work), and that in such cases the dependence on external non-Java software DSSP presents an inconvenient break in the work flow. Furthermore, this new tool also allows direct visualisation of the amino acid sequence of a three-dimensional structure in PDB format with automatically mapped secondary structure elements.

Here, we report on an update of the SBAL software which includes improved parsers to better handle the variety of input file formats, as well as further processing and analysis tools. We also implemented the DSSP algorithm into Java and introduce the new application analysis of secondary structure of proteins (ASSP). ASSP has been embedded with the new version of SBAL, thus eliminating the need for running DSSP as an external programme in a separate step prior to executing SBAL, thus creating a streamlined and user-friendly platform for sequence-structure analyses. ASSP is also available as a stand-alone application to analyse secondary structure of three-dimensional models in PDB file format and can be integrated into other Java applications.

Programme implementation and methods

ASSP working concept

The Dictionary of Protein Secondary Structure (DSSP) was designed by Wolfgang Kabsch and Christian Sander to standardise secondary structure assignment (Kabsch and Sander 1983). DSSP has since become a database of secondary structure assignments and other data for all protein entries in the PDB (Joosten et al. 2011). The DSSP software is available as a web service or stand-alone programme, written in C++ (http://www.cmbi.ru.nl/dssp.html). Despite its popularity and the eminent importance of the concept for structural biology, the algorithm appears not to have been ported to other programming languages.

In the context of our Java programme collection for structural biology and biophysical chemistry (PCSB) (Hofmann and Wlodawer 2002), and specifically for the application SBAL (Wang et al. 2012), a tool for structure-based sequence alignments, we have implemented the DSSP algorithm as described in the original report by Kabsch and Sander (1983) in Java. To avoid naming confusion with the original DSSP, the Java application is called ASSP.

ASSP works in a similar fashion as DSSP, and its Java API also allows for integration with other Java software. As a stand-alone application, the user can read a three-dimensional protein structure in PDB format and the software will analyse all geometric and secondary structure (H-bonding) parameters. Parsing of the PDB file is accomplished by the PDB file class of PCSB, and includes a check for hydrogen atoms being present. If so, the provided hydrogen atoms are kept and used for analysis. If no hydrogen atoms are present in the PDB file, the amide hydrogen atoms will be modelled.

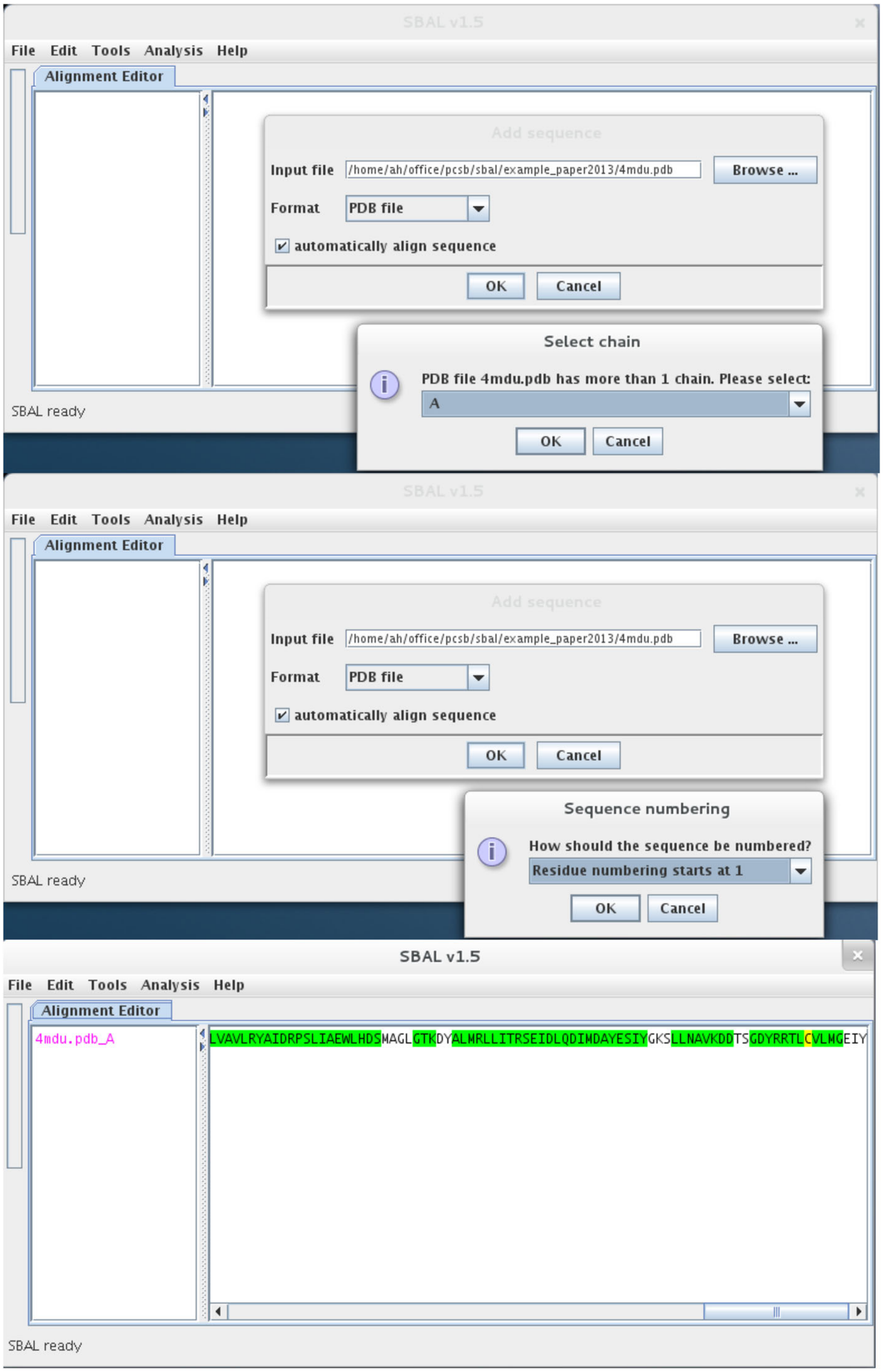
ASSP output

Results are reported either to the terminal or can be saved into an ASCII file. The user can choose to have the results output in DSSP format (to allow for comparison with DSSP, we have used the same format as that delivered by the DSSP web service at http://www.cmbi.ru.nl/hsspsoap/) or in ASSP format. We believe the ASSP output format to be more convenient for visual inspection and manual analysis. Apart from a more generous use of spaces between individual parameters, ASSP reports all residues by their original name as listed in the PDB file. For comparison, DSSP addresses residues by a sequential index.

SBAL

SBAL is a tool to visualise, generate and edit secondary structure-based sequence alignments (Wang et al. 2012). Users can choose to either generate secondary structure-based alignments using SBAL or import sequences with or without secondary structure information from a variety of established input formats, including FASTA, MSF, ClustalW (Larkin et al. 2007), PSIPRED (Bryson et al. 2005), DSSP (Kabsch and Sander 1983) and PDB. When importing information from three-dimensional structures, the secondary structure assignments so far relied on the 'HELIX' and 'SHEET' records in PDB files, or else the user previously needed to pre-process the structure in PDB format with the software DSSP.

ASSP has been embedded into the SBAL application and, when loading a PDB file, the user is prompted with a choice of using the secondary structure information from



Input file /home/ah/office/pcsb/sbal/example_paper2013/4mdu.pdb

Format PDB file

☑ automatically align sequence

Select chain

PDB file 4mdu.pdb has more than 1 chain. Please select: A

Sequence numbering

How should the sequence be numbered? Residue numbering starts at 1

4mdu.pdb_A LVAVLRYAIDRPSLIAEWLHDSMAGLGTKDYALMRLLITRSEIDLQDIMDAYESIYGKSLLNAVKDDTSGDYRRTLCVLMGEIY

Fig. 1 Screen shots of importing PDB file 4mdu (Leow et al. 2014) into SBAL. *Top panel* SBAL recognises the presence of different chains in the file and prompts the user to choose which chain should be imported. *Middle panel* If the starting residue number differs from '1', the user can choose to retain the residue numbering as in the PDB file or re-number from 1. *Bottom panel* Since the PDB file did not have any 'HELIX'/'SHEET' records, SBAL automatically assigns secondary structure using the embedded version of ASSP. Helical structure is indicated in *green*, β-strands are coloured *red*

the ‘HELIX’ and ‘SHEET’ records in the PDB file, or analysing the structure with the built-in ASSP module. Although the import of information from DSSP is still possible, use of the built-in ASSP module provides an extra level of user-friendliness (see Fig. 1).

Availability

Both programmes make use of and extend Java classes previously developed in our laboratory (Hofmann and Wlodawer 2002; Wang et al. 2012). They are available as stand-alone compiled Java applications from the project home page at http://www.structuralchemistry.org/pcsb/. The ASSP API includes methods that enable interfacing with other Java applications and may thus also be useful to developers. The applications and manuals are freely available to academic users. For download, users will be asked for their name, institution and email address. The source code is available from the authors upon request.

Results

Benchmarking of ASSP

To evaluate the secondary structure assignment implementation in ASSP, a dataset of crystal structure files was initially selected using the PISCES server (Wang and Dunbrack 2005), using three criteria: (i) sequence identity below 20 %, (ii) resolution better than 1.6 Å, (iii) R-factor <0.25. This dataset included 2,118 PDB files, from which 100 PDB files (the last 100 of a PDB code-sorted list) were selected as the benchmark set, which comprised over 28,000 residues. The mathematically derived parameters such as hydrogen bonds, angles, etc. reported by ASSP are in excellent agreement with those reported by DSSP. To evaluate the agreement of the secondary structure summary assignment of ASSP with that of DSSP, an in-house script was used to compare the output files of both programmes (see Table 1; example output in Fig. 2). According to DSSP, the benchmark set contains 5,626 residues with no secondary structure assignment, 321 residues in an isolated β-bridge (B), 5,730 residues in an extended strand (E), 1,305 residues in a 3_{10}-helix (G), 9,822 residues in an α-helix (H), 24 residues in a π-helix (I), 2,258 residues in a bend (S), and 2,988 residues in a H-bonded turn (T). ASSP was able to achieve between 85 and 100 % agreement for all cases except for assignments of residues in an isolated β-bridge. Discrepancies in the individual assignments arise when the criteria for more than one type of secondary structure assignment are met and a particular type needs to be assigned based on an hierarchy. This is also true for assignment of the 3_{10}-helical type (G), where ASSP assigns an H-bonded turn (T) to a fraction of $\sim$15 %, since it only allows the assignment of H, G or I for more than three consecutive residues with those types.

Table 1 Distribution of assigned secondary structure types (‘summary’) in the benchmarking dataset analysed with DSSP and ASSP

DSSP assignment	ASSP assignment							
	None	B	E	G	H	I	S	T
None	5,264	150	179	5	23	0	0	5
B	76	156	28	0	2	0	48	11
E	245	49	5,342	2	0	3	73	16
G	24	0	0	909	9	2	160	201
H	3	0	0	547	9,016	104	78	74
I	0	0	0	0	0	24	0	0
S	0	0	16	3	4	0	2,235	0
T	15	5	0	244	81	38	56	2,549

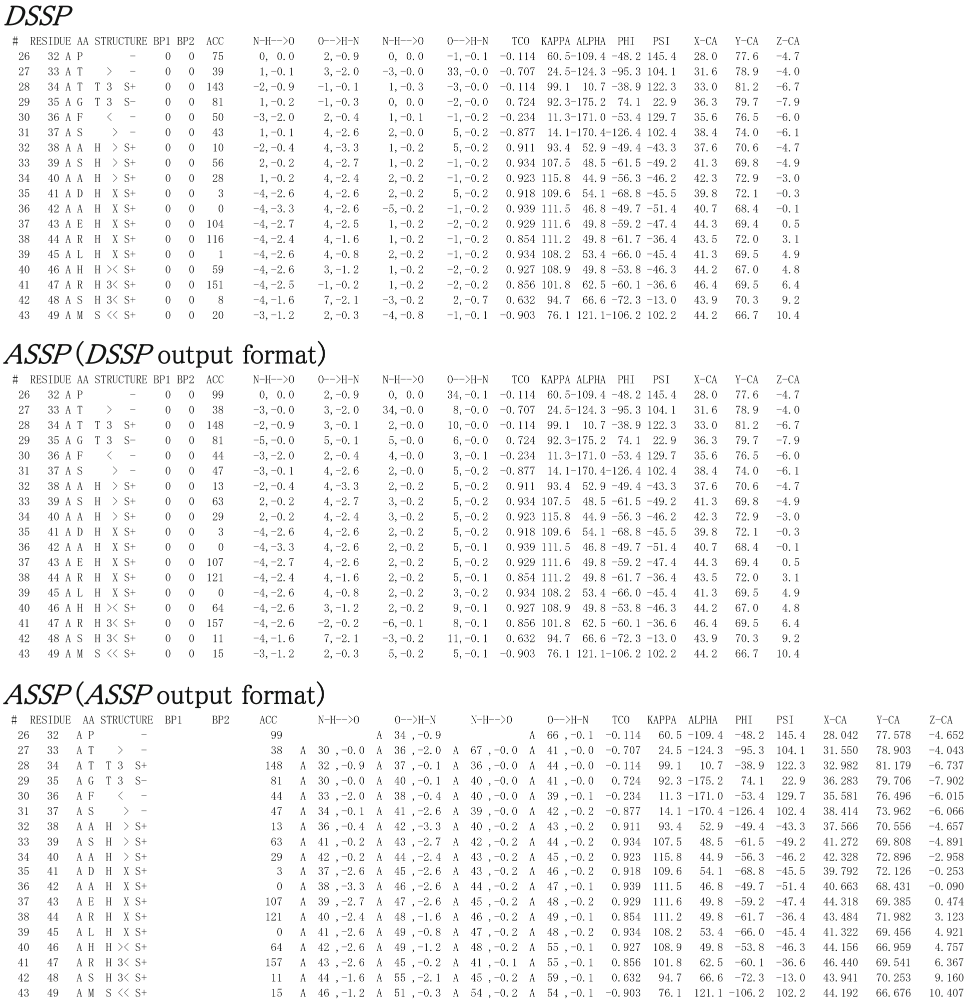
Mapping of secondary structure elements

When mapping secondary structure elements onto an amino acid sequence, SBAL uses three simplified structure types: helical, extended and unstructured. If the information is extracted from fully annotated PDB files, residue ranges listed as ‘HELIX’ are mapped as ‘helical’, residue ranges annotated with ‘SHEET’ are mapped as ‘extended’, and all other residues are mapped as ‘unstructured’. Accordingly, when using either ASSP or DSSP analysis of three-dimensional structures, the three helical secondary structure types (H, I, G) are combined and mapped as ‘helical’, the extended strand (E) is mapped as ‘extended’, and all other types are reported as ‘unstructured’. From the benchmarking exercise, we find an agreement of ASSP and DSSP results of around 90 % or better (see Table 2).

New features of SBAL

With more than 600 downloads since its publication in 2012, SBAL is one of the most popular Java applications of our programme collection. Based on feedback from users, several improvements have been made to address minor glitches, improve parsing of various input file formats, as well as enhance the user experience.

We have embedded an automated secondary structure analysis tool (ASSP) into SBAL that is automatically deployed when a PDB file without ‘HELIX’/‘SHEET’ records is loaded (see Fig. 1); otherwise, the user is prompted to choose which source of secondary structure information to use. Importantly, all three types of helical secondary structure, 3_{10}-helix, α-helix and π-helix, are combined into helical structure at this step.



DSSP

```
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
   26   32 A P         -     0   0   75      0, 0.0     2,-0.9     0, 0.0    -1,-0.1  -0.114  60.5-109.4 -48.2 145.4   28.0   77.6   -4.7
   27   33 A T    >    -     0   0   39      1,-0.1     3,-2.0    -3,-0.0    33,-0.0  -0.707  24.5-124.3 -95.3 104.1   31.6   78.9   -4.0
   28   34 A T  T 3   S+     0   0  143     -2,-0.9    -1,-0.1     1,-0.3    -3,-0.0  -0.114  99.1  10.7 -38.9 122.3   33.0   81.2   -6.7
   29   35 A G  T 3   S-     0   0   81      1,-0.2    -1,-0.3     0, 0.0    -2,-0.0   0.724  92.3-175.2  74.1  22.9   36.3   79.7   -7.9
   30   36 A F    <   -      0   0   50     -3,-2.0     2,-0.4     1,-0.1    -1,-0.2  -0.234  11.3-171.0 -53.4 129.7   35.6   76.5   -6.0
   31   37 A S      >  -     0   0   43      1,-0.1     4,-2.6     2,-0.0     5,-0.2  -0.877  14.1-170.4-126.4 102.4   38.4   74.0   -6.1
   32   38 A A  H  > S+     0   0   10     -2,-0.4     4,-3.3     1,-0.2     5,-0.2   0.911  93.4  52.9 -49.4 -43.3   37.6   70.6   -4.7
   33   39 A S  H  > S+     0   0   56      2,-0.2     4,-2.7     1,-0.2    -1,-0.2   0.934 107.5  48.5 -61.5 -49.2   41.3   69.8   -4.9
   34   40 A A  H  > S+     0   0   28      1,-0.2     4,-2.4     2,-0.2    -1,-0.2   0.923 115.8  44.9 -56.3 -46.2   42.3   72.9   -3.0
   35   41 A D  H  X S+     0   0    3     -4,-2.6     4,-2.6     2,-0.2     5,-0.2   0.918 109.6  54.1 -68.8 -45.5   39.8   72.1   -0.3
   36   42 A A  H  X S+     0   0    0     -4,-3.3     4,-2.6    -5,-0.2    -1,-0.2   0.939 111.5  46.8 -49.7 -51.4   40.7   68.4   -0.1
   37   43 A E  H  X S+     0   0  104     -4,-2.7     4,-2.5     1,-0.2    -2,-0.2   0.929 111.6  49.8 -59.2 -47.4   44.3   69.4    0.5
   38   44 A R  H  X S+     0   0  116     -4,-2.4     4,-1.6     1,-0.2    -1,-0.2   0.854 111.2  49.8 -61.7 -36.4   43.5   72.0    3.1
   39   45 A L  H  X S+     0   0    1     -4,-2.6     4,-0.8     2,-0.2    -1,-0.2   0.934 108.2  53.4 -66.0 -45.4   41.3   69.5    4.9
   40   46 A H  H >< S+     0   0   59     -4,-2.6     3,-1.2     1,-0.2    -2,-0.2   0.927 108.9  49.8 -53.8 -46.3   44.2   67.0    4.8
   41   47 A R  H 3< S+     0   0  151     -4,-2.5    -1,-0.2     1,-0.2    -2,-0.2   0.856 101.8  62.5 -60.1 -36.6   46.4   69.5    6.4
   42   48 A S  H 3< S+     0   0    8     -4,-1.6     7,-2.1    -3,-0.2     2,-0.7   0.632  94.7  66.6 -72.3 -13.0   43.9   70.3    9.2
   43   49 A M  S << S+     0   0   20     -3,-1.2     2,-0.3    -4,-0.8    -1,-0.1  -0.903  76.1 121.1-106.2 102.2   44.2   66.7   10.4
```

ASSP (*DSSP* output format)

```
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
   26   32 A P         -     0   0   99      0, 0.0     2,-0.9     0, 0.0    34,-0.1  -0.114  60.5-109.4 -48.2 145.4   28.0   77.6   -4.7
   27   33 A T    >    -     0   0   38     -3,-0.0     3,-2.0    34,-0.0     8,-0.0  -0.707  24.5-124.3 -95.3 104.1   31.6   78.9   -4.0
   28   34 A T  T 3   S+     0   0  148     -2,-0.9     3,-0.1     2,-0.0    10,-0.0  -0.114  99.1  10.7 -38.9 122.3   33.0   81.2   -6.7
   29   35 A G  T 3   S-     0   0   81     -5,-0.0     5,-0.1     5,-0.0     6,-0.0   0.724  92.3-175.2  74.1  22.9   36.3   79.7   -7.9
   30   36 A F    <   -      0   0   44     -3,-2.0     2,-0.4     4,-0.0     3,-0.1  -0.234  11.3-171.0 -53.4 129.7   35.6   76.5   -6.0
   31   37 A S      >  -     0   0   47     -3,-0.1     4,-2.6     2,-0.0     5,-0.2  -0.877  14.1-170.4-126.4 102.4   38.4   74.0   -6.1
   32   38 A A  H  > S+     0   0   13     -2,-0.4     4,-3.3     2,-0.2     5,-0.2   0.911  93.4  52.9 -49.4 -43.3   37.6   70.6   -4.7
   33   39 A S  H  > S+     0   0   63      2,-0.2     4,-2.7     3,-0.2     5,-0.2   0.934 107.5  48.5 -61.5 -49.2   41.3   69.8   -4.9
   34   40 A A  H  > S+     0   0   29      2,-0.2     4,-2.4     3,-0.2     5,-0.2   0.923 115.8  44.9 -56.3 -46.2   42.3   72.9   -3.0
   35   41 A D  H  X S+     0   0    3     -4,-2.6     4,-2.6     2,-0.2     5,-0.2   0.918 109.6  54.1 -68.8 -45.5   39.8   72.1   -0.3
   36   42 A A  H  X S+     0   0    0     -4,-3.3     4,-2.6     2,-0.2     5,-0.1   0.939 111.5  46.8 -49.7 -51.4   40.7   68.4   -0.1
   37   43 A E  H  X S+     0   0  107     -4,-2.7     4,-2.6     2,-0.2     5,-0.2   0.929 111.6  49.8 -59.2 -47.4   44.3   69.4    0.5
   38   44 A R  H  X S+     0   0  121     -4,-2.4     4,-1.6     2,-0.2     5,-0.1   0.854 111.2  49.8 -61.7 -36.4   43.5   72.0    3.1
   39   45 A L  H  X S+     0   0    0     -4,-2.6     4,-0.8     2,-0.2     3,-0.2   0.934 108.2  53.4 -66.0 -45.4   41.3   69.5    4.9
   40   46 A H  H >< S+     0   0   64     -4,-2.6     3,-1.2     2,-0.2     9,-0.1   0.927 108.9  49.8 -53.8 -46.3   44.2   67.0    4.8
   41   47 A R  H 3< S+     0   0  157     -4,-2.6    -2,-0.2    -6,-0.1     8,-0.1   0.856 101.8  62.5 -60.1 -36.6   46.4   69.5    6.4
   42   48 A S  H 3< S+     0   0   11     -4,-1.6     7,-2.1    -3,-0.2    11,-0.1   0.632  94.7  66.6 -72.3 -13.0   43.9   70.3    9.2
   43   49 A M  S << S+     0   0   15     -3,-1.2     2,-0.3     5,-0.2     5,-0.1  -0.903  76.1 121.1-106.2 102.2   44.2   66.7   10.4
```

ASSP (*ASSP* output format)

```
 #  RESIDUE  AA STRUCTURE  BP1      BP2      ACC        N-H-->O        O-->H-N        N-H-->O        O-->H-N     TCO   KAPPA  ALPHA    PHI     PSI      X-CA     Y-CA     Z-CA
  26   32    A P         -                     99                   A  34 ,-0.9                   A  66 ,-0.1  -0.114    60.5 -109.4   -48.2   145.4    28.042   77.578   -4.652
  27   33    A T    >    -                     38   A  30 ,-0.0   A  36 ,-2.0   A  67 ,-0.0   A  41 ,-0.0  -0.707    24.5 -124.3   -95.3   104.1    31.550   78.903   -4.043
  28   34    A T  T 3   S+                    148   A  32 ,-0.9   A  37 ,-0.1   A  36 ,-0.0   A  44 ,-0.0  -0.114    99.1   10.7   -38.9   122.3    32.982   81.179   -6.737
  29   35    A G  T 3   S-                     81   A  30 ,-0.0   A  40 ,-0.1   A  40 ,-0.0   A  41 ,-0.0   0.724    92.3 -175.2    74.1    22.9    36.283   79.706   -7.902
  30   36    A F    <   -                      44   A  33 ,-2.0   A  38 ,-0.4   A  40 ,-0.0   A  39 ,-0.1  -0.234    11.3 -171.0   -53.4   129.7    35.581   76.496   -6.015
  31   37    A S      >  -                     47   A  34 ,-0.1   A  41 ,-2.6   A  39 ,-0.0   A  42 ,-0.2  -0.877    14.1 -170.4  -126.4   102.4    38.414   73.962   -6.066
  32   38    A A  H  > S+                     13   A  36 ,-0.4   A  42 ,-3.3   A  40 ,-0.2   A  43 ,-0.2   0.911    93.4   52.9   -49.4   -43.3    37.566   70.556   -4.657
  33   39    A S  H  > S+                     63   A  41 ,-0.2   A  43 ,-2.7   A  42 ,-0.2   A  44 ,-0.2   0.934   107.5   48.5   -61.5   -49.2    41.272   69.808   -4.891
  34   40    A A  H  > S+                     29   A  42 ,-0.2   A  44 ,-2.4   A  43 ,-0.2   A  45 ,-0.2   0.923   115.8   44.9   -56.3   -46.2    42.328   72.896   -2.958
  35   41    A D  H  X S+                      3   A  37 ,-2.6   A  45 ,-2.6   A  43 ,-0.2   A  46 ,-0.2   0.918   109.6   54.1   -68.8   -45.5    39.792   72.126   -0.253
  36   42    A A  H  X S+                      0   A  38 ,-3.3   A  46 ,-2.6   A  44 ,-0.2   A  47 ,-0.1   0.939   111.5   46.8   -49.7   -51.4    40.663   68.431   -0.090
  37   43    A E  H  X S+                    107   A  39 ,-2.7   A  47 ,-2.6   A  45 ,-0.2   A  48 ,-0.2   0.929   111.6   49.8   -59.2   -47.4    44.318   69.385    0.474
  38   44    A R  H  X S+                    121   A  40 ,-2.4   A  48 ,-1.6   A  46 ,-0.2   A  49 ,-0.1   0.854   111.2   49.8   -61.7   -36.4    43.484   71.982    3.123
  39   45    A L  H  X S+                      0   A  41 ,-2.6   A  49 ,-0.8   A  47 ,-0.2   A  48 ,-0.2   0.934   108.2   53.4   -66.0   -45.4    41.322   69.456    4.921
  40   46    A H  H >< S+                     64   A  42 ,-2.6   A  49 ,-1.2   A  48 ,-0.2   A  55 ,-0.1   0.927   108.9   49.8   -53.8   -46.3    44.156   66.959    4.757
  41   47    A R  H 3< S+                    157   A  43 ,-2.6   A  45 ,-0.2   A  41 ,-0.1   A  55 ,-0.1   0.856   101.8   62.5   -60.1   -36.6    46.440   69.541    6.367
  42   48    A S  H 3< S+                     11   A  44 ,-1.6   A  55 ,-2.1   A  45 ,-0.2   A  59 ,-0.1   0.632    94.7   66.6   -72.3   -13.0    43.941   70.253    9.160
  43   49    A M  S << S+                     15   A  46 ,-1.2   A  51 ,-0.3   A  54 ,-0.2   A  54 ,-0.1  -0.903    76.1  121.1  -106.2   102.2    44.192   66.676   10.407
```

Fig. 2 Excerpt from the secondary structure analysis of the same PDB file as in Fig. 1 with DSSP (*top section*), and ASSP (*middle and bottom sections*). The type of information reported by ASSP is the same as that of DSSP. The middle section shows a report generated by ASSP mimicking the DSSP output format with relative referencing of residues. For user convenience, ASSP can also produce reports where residues are referenced with their absolute names, i.e. chain identifier and residue number as in the PDB file (*bottom section*)

Table 2 Comparison of simplified ASSP and DSSP secondary structure assignment results obtained with the benchmarking dataset

Sec. structure	DSSP	ASSP	Agreement
Unstructured	11,193	10,204	0.912
Helical	11,151	9,949	0.892
Extended	5,730	5,342	0.932

In the ‘Tools’ section, the calculation of a distance matrix for the current alignment has been added. Pairwise distances for all sequences in the alignment can be calculated based on either the amino acid identity or *p* distances. For the latter, the user has the option to choose *p* distances as used in the EMBOSS programme distmat (Rice et al. 2000), or the programme MEGA (Kumar et al. 2001). The distance matrix is displayed in a separate window and can readily be exported to spreadsheet programmes by a copy–paste operation.

Another addition to the ‘Tools’ section is the option to calculate peptide properties for the currently active sequence in the alignment by just one mouse click. The physical peptide properties calculated include number of amino acids, molecular mass, extinction coefficient at a wavelength of 280 nm, isoelectric point and charge at pH 7 (Hofmann and Wlodawer 2002).

Among the convenience features added are the option to add an annotation to an alignment, a moveable barrier between the ID panel and the sequence panel (to accommodate long sequence titles), and more extensive preference settings that allow changing various colour settings and the number of amino acid residues within one line of the HTML output format. On the project web site (http://www.structuralchemistry.org/pcsb/), a site dedicated to SBAL includes examples of various frequent scenarios with step-by-step instructions.

Conclusions

Amino acid sequence alignments are core to structural biology and alignments with mapped secondary structure elements are essential to inferring structural homology between proteins. SBAL has proven to be a popular integrated Java tool for generation, visualisation and analysis of structure-based sequence alignments based on the download statistics. With development and implementation of a Java tool that conducts geometric/hydrogen bond analysis of three-dimensional protein structures to automatically derive secondary structure assignments, the work flow in generation of structure-based sequence alignments has been significantly simplified, as PDB files can now be directly accessed for sequence alignment purposes, without the need for pre-processing.

The stand-alone ASSP analysis tool as Java application closes a gap in the currently available Java libraries for structural biology and will be useful in any computational structural biology context. The API will further assist in development of future Java classes for structural biology. The current version of ASSP has several built-in methods aiming particularly at modelling applications (e.g. output of information directly usable in structure-based amino acid sequence alignments), and we plan to implement further model analysis tools in the future.

Acknowledgments Research in the laboratory of AH is funded by the National Health and Medical Research Council (APP1002898, APP1044022) and the Australian Research Council (DP140100599, LE120100071, LP120100485).

Conflict of interest The authors declare that they have no conflict of interest regarding this publication.

References

Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33:W36–W38

Cantacessi C, Jex AR, Hall RS, Young ND, Campbell BE, Joachim A, Nolan MJ, Abubucker S, Sternberg PW, Ranganathan S et al (2010) A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing. Nucleic Acids Res 38:e171

Cantacessi C, Seddon JM, Miller TM, Lee CY, Thomas L, Mason L, Willis C, Walker G, Loukas A, Gasser B et al (2013) A genome-wide analysis of annexins from parasitic organisms and their vectors. Sci Reports 3:2893

Cantacessi C, Hofmann A, Campbell BE, Gasser RB (2014) Impact of next-generation technologies on exploring socio-economically important parasites and developing new interventions. In: Inacio J, Cunha M (eds) Molecular diagnostics and high-throughput strategies in veterinary infection biology. Springer, Berlin (in press)

Eidhammer I, Jonassen I, Taylor WR (2000) Structure comparison and structure patterns. J Comput Biol 7:685–716

Hewitson JP, Grainger JR, Maizels RM (2009) Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity. Mol Biochem Parasitol 167:1–11

Hofmann A, Wlodawer A (2002) PCSB–a program collection for structural biology and biophysical chemistry. Bioinformatics 18:209–210

Hubbard TJ, Blundell TL (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. Protein Eng 1:159–171

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S et al (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40:D306–D312

Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G (2011) A series of PDB related databases for everyday needs. Nucleic Acids Res 39:D411–D419

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

Leow CY, Willis C, Osman A, Mason L, Simon A, Gasser RB, Smith B, Jones MK, Hofmann A (2014) Crystal structure and immunological properties of the first annexin from *Schistosoma mansoni*. FEBS J 281:1209–1225

Marchler-Bauer A, Panchenko AR, Ariel N, Bryant SH (2002) Comparison of sequence and structure alignments for protein domains. Proteins 48:439–446

Osman A, Wang CK, Winter A, Loukas A, Tribolet L, Gasser RB, Hofmann A (2012) Hookworm SCP/TAPS protein structure—a key to understanding host-parasite interactions and developing new interventions. Biotechnol Adv 30:652–657

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277

Russell RB, Barton GJ (1994) Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. J Mol Biol 244:332–350

Sauder JM, Arthur JW, Dunbrack RL (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 40:6–22

Wang G, Dunbrack RLJ (2005) PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 33:W94–W98

Wang CK, Broder U, Weeratunga SK, Gasser RB, Loukas A, Hofmann A (2012) SBAL: a practical tool to generate and edit structure-based amino acid sequence alignments. Bioinformatics 28:1026–1027