PAL: A Perl Script for Rapidly Identifying the Active Site of Large Protein Families

Andres A. Larrea.[13] and Pablo A. Larrea[2]

[1] University of Miami Miller School of Medicine Department of Biochemistry and Molecular Biology

[2] Unaffiliated

[3] Corresponding author:
Department of Biochemistry and Molecular Biology
University of Miami Miller School of Medicine
1011 NW 15 St
Miami, FL 33136
ALarrea@med.miami.edu

Keywords: Multiple sequence alignment analysis; active site determination; protein domains

Software URL: http://homepage.mac.com/amlarrea/FileSharing3.html

Word Count: 899

In the post-genomic era, with an ever-increasing amount of sequence information, it is critical to develop tools that assist in sifting through all the data to identify information amidst the noise. The last ten years have seen an explosion in DNA sequence data. As of late March 2007, there were 921 genomic sequence sets in the NCBI database (with 495 completely sequenced genomes) (1). With the continuing rapid expansion in genome data, it is becoming increasingly important to develop tools that assist in the rapid analysis of large numbers of sequences. The ability to parse through large amounts of sequence data is vital as this allows us to identify patterns that can be used to predict functions for unknown open reading frames (ORFs).

The two main steps in determining the functional sites in proteins are: (1) aligning a large set of homologous sequences and (2) analyzing this large file to identify highly conserved residues. A great deal of software exists for generating a quality alignment (*e.g.* Clustal (2-4), Muscle (5, 6), T-Coffee (7-9), etc). Although, historically computation time has been a limiting factor in aligning large files, with improvements in computer technology as well as in the algorithms, some of these programs are able to generate quality multiple sequence alignment (MSA) files very quickly. Generating an MSA file is, however, only the first step in analyzing proteins. The real goal is to allow genetic variation and natural selection to point us in the direction of important amino acid residues as evident in the degree of conservation and pattern of inheritance in phylogenetic trees(10). Unfortunately there is no software able to quickly analyze large multiple sequence alignment files. Most of the software that exists for post-alignment analysis is Graphical User Interface (GUI) based (JalView (11), SeaView (12), ClustalX (4)). Although GUIs have several advantages, looking at large amounts of data and identifying important residues, is not one of them.

Once an MSA file has been created, it is important to be able to sift through it quickly to identify conserved residues that often mark functionally important protein features like enzyme active sites. To this end, we have written PAL, a small highly customizable Perl script. While the inputs and outputs are straight-forward (**Figure 1**) they can be looked at in a variety of ways to solve several problems faced by investigators as they analyze multiple sequence alignments:

1. Generating a high quality alignment is an iterative process. It is important to include a large number of sequences in a multiple sequence alignment, but sequences that are too divergent can introduce noise that obscures gene function, while identical sequences offer no predictive benefit. By analyzing the *nearest.txt* **(Figure 1G)** file, users can quickly identify which sequences should be removed before the next alignment cycle.

2. While enzyme specificity is evolutionarily flexible, novel chemistry is far more difficult to evolve and active site residues are predicted to be highly conserved (13). Once the number of sequences being used has been trimmed down and the alignment is finalized, the files *consensus.txt* **(Figure 1D)** and *nearconsensus.txt* **(Figure 1E)** can be used to predict putative enzyme active sites. These sites can then be verified experimentally using site-directed mutagenesis and enzymatic analysis approaches.

3. Although enzyme chemistry is difficult to evolve and reflected in highly conserved regions in a protein sequence, substrate specificity and protein-protein interaction surfaces are often sites of divergence between homologs. Examination of the file *letters.txt* (**Figure 1F)**, output by PAL, can be used to identify highly divergent regions that may be involved in specificity.

4. With very large protein families, it can be helpful to divide the sequences into smaller sub-families for analysis. By using the *group.analysis* option in the inputs, users can choose to

divide the full family into sub-families with different amounts of homology. Each sub-family is then individually analyzed for conservation and divergence by PAL.

The only input requirement is an alignment in fasta format. Because of this, PAL can use any alignment software as well as manually edited alignments. What makes PAL so powerful is that it is able analyze very large alignment files in a short period of time (**Figure 2**).

The blessing of an ever-increasing amount of genomic data is also a curse that leads to large MSA files that are difficult to analyze using traditional graphical approaches. In this paper we present a Perl script that assists in the rapid identification of important residues in large protein families. With this software we have been able to identify a putative active site for a large exonuclease family containing over 300 members in less than 2 minutes. The sample outputs shown in **Figure 1** have formed the basis of ongoing work to understand the mechanism of the poorly understood ExoVII nuclease family of enzymes (Larrea A *et. al.*, unpublished data).

Software requirements and availability:

All software described here is open source and freely available upon request. Although all benchmarking has been done on a Mac, much of the programming was done on a Windows platform PC. Since the software is written it Perl it should run the same across all platforms.

Acknowledgements:

This work was supported by a pre-doctoral fellowship from the NIH (F31-GM70395 to AAL). Many thanks to Arun Malhotra and Richard Myers for their invaluable help during the beta testing of the software and during the preparation of the manuscript.

Competing Interest:

The authors have no competing interests.

1.**Cummings, L., L. Riley, L. Black, A. Souvorov, S. Resenchuk, I. Dondoshansky, T. Tatusova.**2002. Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes. FEMS Microbiol Lett 2:133-8.

2.**Higgins, D.G., P.M. Sharp.**1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 1:237-44.

3.**Thompson, J.D., D.G. Higgins, T.J. Gibson.**1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-80.

4.**Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins.**1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876-82.

5.**Edgar, R.C..**2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 113.

6.**Edgar, R.C..**2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 5:1792-7.

7.**O'Sullivan, O., K. Suhre, C. Abergel, D.G. Higgins, C. Notredame.**2004. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol 2:385-95.

8.**Poirot, O., K. Suhre, C. Abergel, E. O'Toole, C. Notredame.**2004. 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. Nucleic Acids Res Web Server issue:W37-40.

9.**Poirot, O., E. O'Toole, C. Notredame.**2003. Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. Nucleic Acids Res 13:3503-6.

10.**Livingstone, C.D., G.J. Barton.**1996. Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol 497-512.

11.**Clamp, M., J. Cuff, S.M. Searle, G.J. Barton.**2004. The Jalview Java alignment editor. Bioinformatics 3:426-7.

12.**Galtier, N., M. Gouy, C. Gautier.**1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 6:543-8.

13.**Petsko, G.A..**2000. Enzyme evolution. Design by necessity. Nature 6770:606-7.

Figure 1. PAL inputs and outputs

(A) Table of inputs and outputs for PAL. **(B-G)** Sample outputs from PAL analysis of an alignment with 345 sequences homologous to *E. coli* XseA (gi:16130434). **(B)** Portion of *names.txt*. Section shown only includes sequences numbered 181 to 191. This file references sequence numbers to sequence identifiers. **(C)** Portion of *lettercnt.txt*. This file shows amino acid or gap occurence as a function of position along the alignment. For example, the top line indicates that 285 homologs out of 345 contain a gap at position 181. **(D)** Complete output from *consensus.txt*. This file lists residues that are 100% conserved among all 345 sequences. These residues are likely to form the active site and work is underway to verify this prediction experimentally. **(E)** Complete output from *nearconsensus.txt*. This file lists positions that only include amino acids defined as similar in the *near.match* file. **(F)** Portion of *letters.txt*. This is an alternate form of the data shown in *lettercnt.txt*. In this case each position along the alignment is listed with the residues found at that position in decreasing abundance. For position 181 in the alignment, a gap is seen most often, and Met is the next most abundant residue. The region shown (position 181-190 in the sequence) identifies a putative loop missing in most XseA homologs. **(G)** Portion of output from *nearest.txt*. Matrix of each sequence, with the pairwise value to itself, and to every other sequence. Pairwise values are obtained using the matrix chosen (in this case *blosum62.mtx*). Top ten scores are shown.

Figure 2. Benchmarking PAL

Benchmarking of PAL on G4 PPC and Intel C2D processors of alignments generated by ClustalW with either 145 or 345 sequences.

A

Filename	Type	Description
		INPUTS
XXX.fasta	alignment	aligned protein sequences in FASTA format
XXX.mtx	matrix	alignment matrix to be used. blosum62.mtx and identity.mtx are two supplied but any can be generated.
near.match	text	list of amino acids that are to be treated as equivalent
group.analysis	text	threshold of group definition to be used
		OUTPUTS
names.txt	list	list of the names of all the sequences read (B)
lettercnt.txt	matrix	matrix of position vs. amino acid. expanded form letters.txt (C)
consensus.txt	list	list of residues that are 100% conserved (D)
nearconsensus.txt	list	list of residues that "similar" as defined by near.match (E)
letters.txt	list	list of position vs. amino acid (F)
nearest.txt	matrix	matrix of top ten pairwise scores (G). In each case the sequence is first evaluated against itself and then against every other sequence. The top 10 values are listed first sequence number and then pairwise score.

B

Num	Description
181	tr|Q83C85|Q83C85_COXBU
182	tr|Q3K7C5|Q3K7C5_PSEPF
183	tr|Q4K6V9|Q4K6V9_PSEF5
184	tr|Q48LY3|Q48LY3_PSE14
185	tr|Q4ZX12|Q4ZX12_PSEU2
186	sp|Q886X9|EX7L_PSESM
187	sp|Q88P26|EX7L_PSEPK
188	tr|Q2X8X3|Q2X8X3_PSEPU
189	sp|Q9HXL8|EX7L_PSEAE
190	tr|Q4IXK7|Q4IXK7_AZOVI
191	tr|Q44IK4|Q44IK4_CHRSL

C

	-	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
181	285	6		2	2		2		3	2	4	13	1	6	3	1	5	7	3		
182	283	15		2	5	3	2	5	2	2	5	4	1	7	2	3	2		2		
183	281	4		4	3	3		1		3	2	3	5	5		3	16	4	6	2	
184	267	6	6	11	7	2	3		2	2	1	14	2	4	2	4	6	4	2		
185	248	7	1	2	5	1	3	1	3	1	10	25	2	7	2	3	14	9	1		
186	236	11		7	4	2	4		12	6	8	15	2	9	4	4	10	8	2	1	
187	223	9	2	8	6	13	6		2	15	11	17	5	5		8	6	7	2		
188	214	11		19	8	6	5		4	2	21	13	2	4	2	5	20	6	3		
189	215	11	1	18	9	8	5	2	2	5	7	3	3	17	3	8	17	7	2	2	
190	208	8		5	6	18	5	4	1	5	8	6	10	16	4	6	14	10	11		

D

From	To	Pattern
453	456	RGGG
485	486	GH
489	489	D
494	494	D
504	504	P
730	730	-

E

At	Near consensus
461	L M I
483	A G S
484	V I L
492	I L V M
503	T S
505	T S

F

Pos	Letter frequency
181	- M T P A S L I Q V K E G D R N
182	- A P H L E M F R G D I Q K S V N
183	- S V N P A T D F R K M E L W H
184	- M D E C A S R P T G I Q K F N V L
185	- M S L T A P E R I G N D Q C H F K V
186	- M I A S P L T D K Q R G E N F V W
187	- M K F L A D R T G S E P N C I V
188	- L S D M A E F T G R P I V K Q N
189	- D P S A E R F L T K G N M Q I H V W C
190	- F P S V T N L A E M R K D G H Q I

G

Struc	Total	With	Sum	With	Sum	With	Sum	With	Sum	With	Sum	With	Sum	With	Sum	With	Sum	With	Sum	With	Sum	With	Sum
181	2045	169	939	167	926	168	874	192	868	136	863	146	862	135	858	164	857	163	852	144	849		
182	2260	183	2129	184	2006	186	2002	185	1993	187	1915	188	1904	189	1901	190	1668	168	1059	164	1027		
183	2260	182	2129	186	2013	184	1995	185	1987	189	1926	187	1925	188	1921	190	1688	168	1066	164	1034		
184	2253	185	2202	186	2190	182	2006	183	1995	187	1874	188	1865	189	1846	190	1615	164	1046	168	1033		
185	2277	184	2202	186	2198	182	1993	183	1987	188	1870	187	1856	189	1836	190	1618	164	1040	168	1030		
186	2257	185	2198	184	2190	183	2013	182	2002	187	1852	189	1849	188	1843	190	1606	168	1033	164	1028		
187	2253	188	2234	183	1925	182	1915	184	1874	189	1861	185	1856	186	1852	190	1677	168	1060	164	1030		
188	2275	187	2234	183	1921	182	1904	185	1870	184	1865	189	1854	186	1843	190	1691	168	1042	164	1020		
189	2250	183	1926	182	1901	187	1861	188	1854	186	1849	184	1846	185	1836	190	1759	167	1065	168	1060		
190	2449	189	1759	188	1691	183	1688	187	1677	182	1668	185	1618	184	1615	186	1606	164	1070	168	1045