Georgia Journal of Science

Volume 80 *No. 2 Scholarly Contributions from the Membership and Others* Article 12

2022

USING THE EMBL-EBI CLUSTAL OMEGA TOOL TO CALCULATE DIVERSITY OF HEAVY CHAIN PHAGE-DISPLAY LIBRARIES

Michael Bodri
University of North Georgia, michael.bodri@ung.edu

Shane A. Webb
University of North Georgia, shane.webb@ung.edu

USING THE EMBL-EBI CLUSTAL OMEGA TOOL TO CALCULATE DIVERSITY OF HEAVY CHAIN PHAGE-DISPLAY LIBRARIES

Acknowledgements

The authors would like to thank the Pearson Pond Ranch and Llama Co. for allowing llama blood collection. Timothy Kennell, Jr. performed some of the laboratory procedures.

USING THE EMBL-EBI CLUSTAL OMEGA TOOL TO CALCULATE DIVERSITY OF HEAVY CHAIN PHAGE-DISPLAY LIBRARIES

Michael S. Bodri
Shane A. Webb

Department of Biology
University of North Georgia
Dahlonega, GA 30597

Corresponding author: Shane A. Webb
shane.webb@ung.edu

ABSTRACT

Here we show that traditional Sanger sequencing combined with analysis tools available from the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), specifically EMBOSS Transeq and Clustal Omega, is extremely effective in the analysis of naïve phage display antibody libraries for the determination of library size and diversity. The free tools are easy to use and require little manipulation of reads by hand, allowing analysis to be performed on a standard personal computer. Utilization of this technique has applicability to researchers with limited access to deep sequencing. The primary drawback to this analysis methodology is that antibodies with particular molecular or binding properties, desirable or otherwise, are not identifiable by their sequences.

***Keywords*:** Clustal Omega, EMBOSS Transeq, HCAb, Fab, heavy-chain antibody, library diversity, library size, phage display, scFv, V_HH

INTRODUCTION

Phage display libraries have become increasingly useful for research and therapeutic purposes that depend upon protein-protein interactions (e.g., enzyme inhibition, receptor activity [agonists and antagonists], drug and vaccine development) [reviewed in 1]. In short, viruses are simple in structure, and can be modified to express other (i.e., non-viral) proteins externally. If these proteins are highly variable, amongst the variation there might exist variants that are biologically useful. Antibody DNA sequences, because of the presence of great variation (hypervariable regions), are therefore an obvious target for library construction.

Filamentous bacteriophage (viruses), typically M13, f1 or fd, with foreign coding sequences spliced into the genes for major coat protein pVIII, as well as the minor coat proteins pIII or pVI, and subsequently displayed on the surface of the virion, constitutes a phage-display library [2]. Phage antibody (Ab) libraries, which can consist of billions of individual clones, utilize DNA coding for variable antigen-binding domains (V regions) of antibody molecules for the coat protein-peptide construct [3]. A neutral linker can be used to artificially join the V region of the Ab heavy chain (V_H) with the V region of the light chain (V_L) for expression as a stabilized single polypeptide fragment, or single-chain fragment variable (scFv). The production of scFvs requires separate amplification of both

V_H and V_L components which are then assembled into a single Fv product using a linker, a critical but laborious process [4]. Multimer formation may result from linkers too short while proteolysis or weak domain association may result from linkers that are too long [4]. Alternative methods that avoid these limitations are desirable.

In 1993 it was discovered that the Camelidae had an additional Ab class that consisted solely of H-chains; they lacked an L-chain that in other vertebrates randomly pairs with an H-chain forming heteromeric molecules that diversify the antigen-binding repertoire [5]. These heavy-chain antibodies (HCAbs) have variable domains (V_HHs) that differ significantly from conventional Ab H-chain V_Hs and form a monophyletic cluster within the V_H subgroup C [5]. Using camelid HCAbs in library construction would therefore provide a simpler approach if sequence diversity is sufficiently large (i.e., contain variants that are biologically useful).

Construction of a large library is important to insure high-affinity fragments as well as to determine the success rate of selection of phage against a large repertoire of different antigens [6]. Naïve libraries are frequently used as single pot resources from which high-affinity antibodies with a high success rate of binding activities against a large set of antigens can be isolated. Of challenge is a standardized method for the determination of diversity within a phage display antibody library. Diversity is also called the complexity or size of the library; in theory it is equal to or less than the number of unique clones in the library. The diversity of natural (versus synthetic) antibody libraries is due to contributions from the hypervariable complementarity-determining regions (CDRs) presented on multiple frameworks, and the random combinations of L and H-chains [7, 8]. Diversity of a naïve antibody library (i.e., one from an immune system not exposed to an antigen challenge) is typically performed by counting the initial number of transformed clones before the library is amplified. The number of transformants per transformation is determined after serial dilution. To reduce clonal amplification, several thousand transformations can be performed separately until quality control is completed.

Some of the earliest measures of diversity were based on scFv amplification from individual colonies of transformants with subsequent endonuclease digestion and examination of the restriction patterns [9]. Now, high quality libraries are typically based on transformation efficiency (size) and sequence sampling of the library demonstrating molecular complexity [7, 10], although apparent diversity is often based solely on initial library size out of a representative portion of the library [11, 12]. Library size is often limited by the transformation efficiency of the library DNA in bacteria [8]. Validation of diversity has been based on as few as 15-20 unique sequences of individual clones with alignment of amino acid and nucleotide sequences of the variable domains of scFv [13]. With large datasets generated by high-throughput sequencing, diversity has also been assessed by nonredundant capture-recapture analysis [7]. Functional library size is lower due to removal of irrelevant clones (frameshift, stop codon or deletion products [11]) or instabilities that lead to insert loss [12].

In this paper we present a simple and informative method that relies on cost-efficient dideoxy termination sequencing of a small number of transformants expressing V_HHs, and uses free and publicly available tools to determine library size and perform diversity analysis. The method could also be applied to other phage libraries, such as scFvs.

MATERIALS & METHODS

Construction of a naive llama phage-display immune library was created utilizing blood collected from four, non-immunized llamas (*Lama glama*) from Pearson Pond Ranch and Llama Co. (Ellijay, GA). Briefly, lymphocytes were isolated (Lymphoprep Tube, greiner bio-one, Longwood, FL) from anti-coagulated blood collected from seven llamas. Total RNA was extracted (Promega PureYield Midiprep System, Promega Corporation, Madison, WI) from an aliquot of 5 x 10^7 lysed cells. Isolation of poly (A) RNA was achieved through biotinylated, oligo(dT) primer hybridization capture and washing in high stringency conditions using streptavidin paramagnetic particles (PolyATtract mRNA Isolation Systems, Promega Corporation, Madison, WI). Eluted mRNA was reverse transcribed into cDNA prior to amplification (ImProm-II reverse Transcription System, Promega Corporation, Madison, WI) (Table 1).

Table 1. Ingredients and thermal profile for the production of cDNA from mRNA collected from llama blood.

Ingredients	
llama mRNA (25 ng)	4 µl
random hexamer RNA	1 µl
5x reaction buffer	4 µl
RNasin (400 U/ µl)	0.5 µl
reverse transcriptase	1 µl
equimolar dNTPs	1 µl
25 mM $MgCl_2$	4.8 µl
dH_2O	3.7 µl
TOTAL	20 µl

Thermal profile	
25^{0}C	5 min
42^{0}C	60 min
70^{0}C	15 min
70^{0}C	15 min

V_HHs gene fragments were amplified by PCR of the cDNA template (Table 2). Fv sequences, the smallest fragment made from enzymatic cleavage of V_HHs (IgG and IgM class Abs) can be selectively amplified from the cDNA by PCR using specific, complementary oligonucleotide primers. Primers complementary to the 5' end recognize either the leader sequence or the conserved parts of the first framework region [14, 15]. Primers that recognize the 3' end are complementary to the constant region. Two primer sets derived from dromedary camel (*Camelus dromedarius*) were used: 5'-GTCCTGGCTGCTCTTCTACAAGG-3' (llama1) annealing at the leader sequence and 5'-GGTACGTGCTGTTGAACTGTTCC-3' (llama2) annealing at the CH2 exon of the heavy chain [14], and; 5'-GTCCTGGCTCTCTTCTACAAGG-3' (Altllama01) [15] annealing at the leader sequence with llama2. The llama1 and llama2 primers have 100% sequence identity with the *Lama glama* genome (confirmed by BLAST search of Genbank). PCR yielded two products, one of ~952 bp and a second of ~ 709 bp. The smaller product, representing V_HH-CH2 without the CH1 exon was excised and recovered from the gel (Promega Wizard SV Gel and PCR Clean Up, Promega Corporation, Madison, WI).

Table 2. Ingredients and thermal profile for the amplification (PCR) of HCAb products from llama blood cDNA.

Ingredients	
cDNA	10 ul
forward primer (10 uM)	10 ul
reverse primer (10 uM)	10 ul
5x reaction buffer	18 ul
RNasin (400 U/ul)	0.5 ul
Taq DNApol	0.5 ul
equimolar dNTPs	1 ul
25 mM $MgCl_2$	5.6 ul
dH_2O	44.9 ul
TOTAL	100 ul

	Thermal profile	
	95^{0}C	2 min
25x	95^{0}C	1 min
	65.5^{0}C	1 min
	72^{0}C	2 min
	72^{0}C	5 min
	4^{0}C	hold

Restriction enzyme sites *Sfi*I and *Not*I were added at either end of the amplicon by nested PCR. Again, two sets of primers, complementary to dromedary (Nestfor, Nestback and AltNfor, AltNback) were used to amplify both products from the first PCR reaction: 5'-GGACTAGTGCGGCCGCGTGAGGAGACGGTGACCTG-3' (Nestfor) and 5'-CATGCCATGACTCGCGGCCCAGCCGGCCATGGCCGAKGTSCAGCT-3' (Nestback) containing *Not*I and *Sfi*I restriction enzyme sites respectively [16], and; 5'-GGACTAGTGCGGCCGCTGCAGACGGTGACCTGGGT-3' (AltNfor) and 5'-GCGGCCCAGCCGGCCGAGTCTGGAGGAGG-3' (AltNback) [15], preserving the *Not*I site in AltNfor, and with the *Pst*I site in AltNback replaced by the *Sfi*I site. This resulted in four products of approximately 475 bp (Table 3). While the product is smaller, these nested primers target the V_HHs domain. Identical restriction sites are present in the pCANTAB 5 E phagemid vector (Amersham Biosciences, Piscataway, NJ). PCR products were restriction enzyme digested and a ligation reaction performed to create vectors containing the antibody V_HH gene followed by electroporation-mediated transformation of competent *E. coli* TG1 cells (Amersham Biosciences, Piscataway, NJ). Cells were harvested and frozen, producing four libraries that ranged in size from between 2.49 x 10^7 and 4.3 x 10^{18} clones for each of the original PCR products.

Table 3. Experimental design with respect to the successive sets of PCR primers used in this study.

phage display library	First PCR primers	Nested PCR primers
llama	llama1 & llama2	Nestfor & Nestback
Alt llama	Altllama01 & llama2	Nestfor & Nestback
llama Alt	llama1 & llama2	AltNfor & AltNback
Alt llama Alt	Altllama01 & llama2	AltNfor & AltNback

Library heterogeneity was checked by selecting a minimum of 50 individual transformed colonies per library for analysis by single-pass Sanger sequencing (Beckman Coulter Genomics, Danvers, MA). Plasmids and PCR amplicons were purified using SPRI Technology (Beckman Coulter Genomics, Danvers, MA). DNA sequencing was performed using BigDyeTerminator v3.1, with post reaction dye terminator removal using Agencourt CleanSEQ (Beckman Coulter Genomics, Danvers, MA). Sequence delineation

was performed on a 96-capillary ABI PRISM 3730xl with base calling and data compilation. Passing reads, based on the National Human Genome Research Institute's standard definition, were results that yielded a minimum of 100 PHRED 20 bases.

Sequence analyses were performed with the tools available from the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) [17, 18]. The DNA sequences were translated, all 3 forward frames, with EMBOSS Transeq [19]. All out-of-frame protein sequences were removed by hand. The remaining forward frames were compared with the immunoglobulin gamma 1b heavy chain of the camelid *Camelus bactrianus* (GenBank HE653094.1) to check protein reads. All reads were confirmed to contain the variable region including the complementary determining regions (CDRs) 1, 2 and 3. The protein sequences were then aligned and the percent identity matrix was calculated with Clustal Omega of the EMBL-EBI [20]. Pairwise sequence comparisons with high identity scores (>98%) were checked by hand to remove sequences that differed only in length at the ends, presumably due to sequencing-read artifacts near primers. Final confirmation of alignment and diversity was assessed with a Kabat-Wu plot of the aligned sequences [21, 22]. Kabat-Wu variability is calculated on a per amino acid site basis as (sample size x number of different a.a.'s occurring)/(count of maximally occurring a.a.). Higher scores indicate more variability.

RESULTS

Total library sizes for the four primer sets varied from 2.49×10^7 to 4.3×10^{18}, with the proportion of those sequences that were unique ranging from 40-94%. Unique library sizes therefore ranged from 2.29×10^7 to 1.72×10^{18} (Table 4). The 50 to 55 translated sequences per library had an aligned length of 185 amino acids, with the majority of length variation due to indels in the third CDR (L3 hypervariable) region. The four libraries, with redundant protein sequences removed, ranged in average sequence identity from 73.4-78.5% (Table 4). Variability along the length of the protein, as assessed with a Kabat-Wu plot, revealed the three hypervariable regions characteristic of the heavy-chain antibody (Figure 1).

Table 4. Phage display library sizes for all four primer sets used in this study, and sequence identity of unique protein sequences (duplicates removed). See Methods for primer combinations.

phage display library	total library size	% unique protein sequences	unique library size	average sequence identity
llama	3.70×10^{13}	94	3.48×10^{13}	77.9 (6.2)
Alt llama	2.49×10^7	92	2.29×10^7	75.3 (10.0)
llama Alt	7.06×10^{14}	67.3	4.88×10^{14}	73.4 (9.4)
Alt llama Alt	4.3×10^{18}	40	1.72×10^{18}	78.5 (10.5)

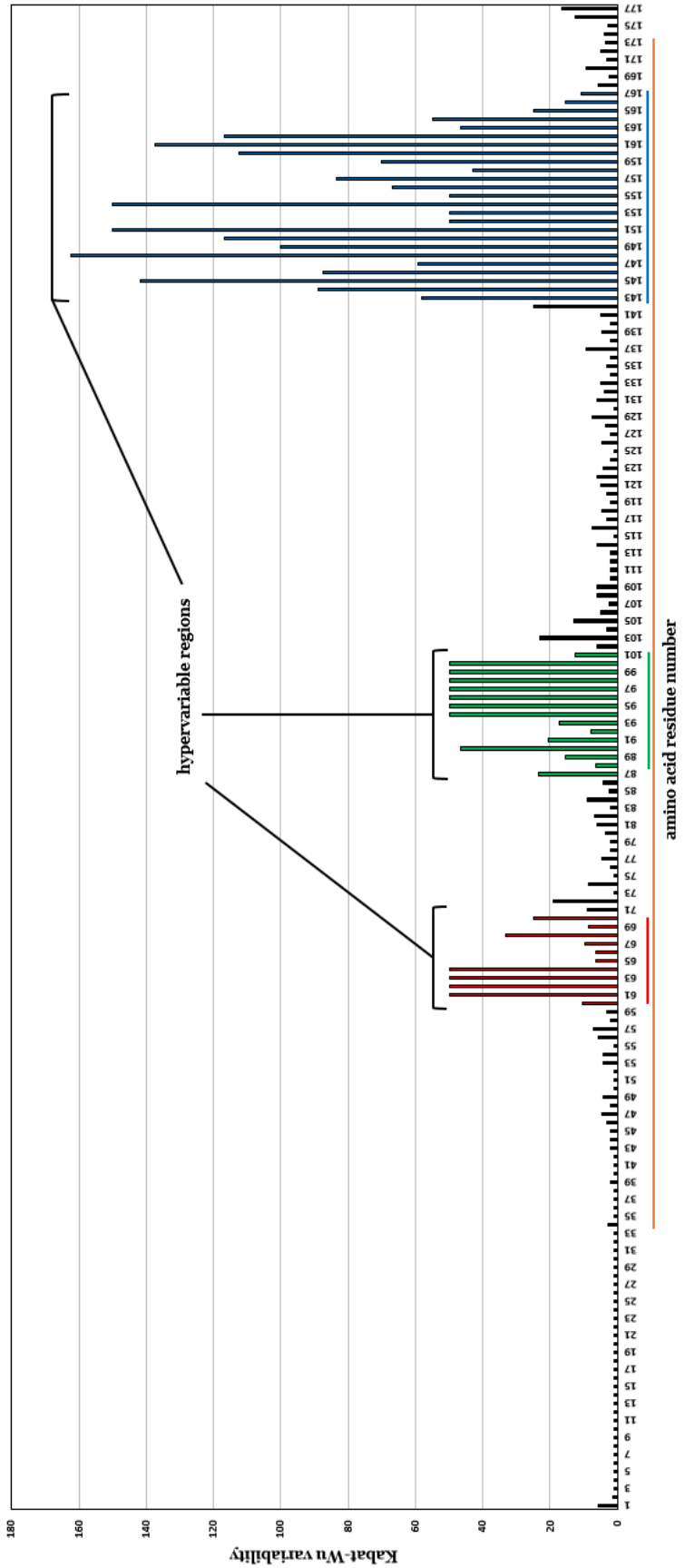
hypervariable regions
amino acid residue number
Kabat-Wu variability
0
20
40
60
80
100
120
140
160
180

Figure 1. Kabat-Wu plot for the llama phage display library based on 50 translated sequences. The variable (V) region is indicated in orange, and the first, second, and third hypervariable (CDR) regions are indicated in red, green, and blue, respectively (following [23]). Plot generated with Microsoft Excel.

DISCUSSION

As previously mentioned, most methods of Ab library size estimation depend in large or entire part upon transformation efficiency [8, 9, 10, 11, 12]. While some studies depend upon sequencing of colonies, this is usually done without assessment of protein diversity [7, 10]. Studies that best assess library size depend upon high throughput sequencing methodologies that are costly, and yield very large numbers of sequences which require a complicated analytical pipeline.

We have demonstrated here that traditional, low-cost, Sanger sequencing combined with the correct tools available from EMBL-EBI, specifically EMBOSS Transeq and Clustal Omega, can be extremely effective in the analysis of naïve phage display antibody libraries for the determination of library size and protein diversity. These free software tools are easy to use and require little manipulation of reads by hand and allow analysis to be performed on a standard personal computer. Utilization of this technique has applicability to researchers with limited access to deep sequencing due to the expense of purchasing and operating the systems, the cost of sequencing, or the complexity of downstream processing necessary for analysis of high-throughput (i.e., next generation) datasets. The primary drawback to this analysis methodology is that antibodies with particular molecular or binding properties, desirable or otherwise, are not identifiable by their sequences.

While there were differences in library size, percent sequence identity of the four combinations of primers suggests any pairwise primer combination is viable for establishing library utility. Curiously, our largest library had the lowest frequency of unique sequences.

Competing interests: The authors declare no competing interests.

Acknowledgements: The authors would like to thank the Pearson Pond Ranch and Llama Co. for allowing llama blood collection. Timothy Kennell, Jr. performed some of the laboratory procedures.

REFERENCES

Papers of special note have been marked as: * of interest

1. Pande, J., M.M. Szewczyk, and A.K. Grover. 2010. Phage display: Concept, innovations, applications and future. Biotech. Adv. 28, 849-858.
2. Petrenko, V.A. and V.J. Vodyanoy. 2003. Phage display for the detection of biological threat agents. J. Microbiol. Methods, 53, 253-262.
3. Rader, C. and C.F. Barbas. 1997. Phage display of combinatorial antibody libraries. *Curr. Opin. Biotechnol.* 8, 503-508.

4. Sotelo, P.H., N. Collazo, R. Zuñiga, M. Gutierrez-Gonzalez, D. Catalan, C. Hager Ribeiro, J. Carlos Aguillon, and M. Carmen Molina. 2012. An efficient method for variable region assembly in the construction of scFv phafge display libraries using independent strand amplification. mAbs, 4, 542-550.
5. Nguyen, V.K., C. Su, and S. Muyldermans. 2002. Heavy-chain antibodies in Camelidae; a case of evolutionary innovation. Immunogenetics, 54, 39-47.
6. Ahmad, Z.A., S.K. Yeap, A.M. Ali, W.Y. Ho, N.B.M. Alitheen, and M. Hamid. 2012. scFv antibody: Principles and clinical application. Clin. Dev. Immunol., article ID 980250, 15 pages. doi:10.1155/2012/980250.
*An overview of single-chain fragment variable antigen binding antibody. The paper stress application as biomarkers and immunotoxins, for therapeutic gene delivery, and as anticancer intrabodies.
7. Glanville, J., W. Zhai, J. Berka, D. Telman, G. Huerta, G.R. Mehta, I. Ni, L. Mei, P.D. Sundar, G.M.R. Day, D. Cox, A. Rajpal, and J. Pons. 2009. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. PNAS, 106, 20216-20221.
* This paper reports on an approach to determining diversity that relies on capture-recapture approaches after high-throughput pyrosequencing. It demonstrates the obstacles facing library developers that may have to rely only on Sanger-based sequencing for the determination of library size and diversity.
8. Lee, C.V., W.C. Liang, M.S. Dennis, C. Eigenbrot, S.S. Sidhu, and G. Fuh. 2004. High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold. J Mol Biol., 340, 1073-1093.
9. Marks, J.D., H.R. Hoogenboom, T.P. Bonnert, J. McCafferty, A.D. Griffiths, and G. Winter. 1991. By-passing immunization: Human antibodies from V-gene libraries displayed on phage. J. Mol. Biol., 222, 581-597.
10. Chen, W., Z. Zhu, Y. Feng, X. Xiao, and D. Dimitrov. 2008. Construction of a large phage-displayed antibody domain library with a scaffold based on a newly identified highly soluble, stable heavy chain variable domain. J. Mol. Biol., 382, 779-789.
11. Knappik, A., L. Ge, A. Honegger, P. Pack, M. Fischer, G. Wellnhofer, A. Hoess, J. Wolle, A. Pluckthun, and B. Virnekas. 2000. Fully synthesized human combinatorial antibody libraries (HuCAL) based on modular concensus frameworks and CDRs randomized with trinucleotides. J. Mol. Biol., 296, 57-86.
12. Willats, W.G.T. 2002. Phage display: practicalities and prospects. Plant Mol. Biol., 50, 837-854.
13. Shavsavarian, M.A., D. Le Minoux, K.M. Matti, S. Kaveri, S. Lacroix-Desmazes, D. Boquet, A. Friboulet, B. Avalle, and S. Padiolleau-Lefevre. 2014. Exploitation of rolling circle amplification for the construction of large phage-display antibody libraries. J. Immunol. Methods, 407, 26-34.
14. Conrath, K.E., M. Lauwereys, M. Galleni, A. Matagne, J.M. Frere, J. Kinne, L. Wyns, and S. Muyldermans. 2001. β-lactamase inhibitors derived from single domain antibody fragments elicited in the camelid. Antimicrob. Agents Chemother., 45, 2807-2812.

15. Saerens, D., J. Kinne, E. Bosmans, U. Wernery, S. Muyldermans, and K. Conrath. 2004. Single domain antibodies derived from dromedary lymph node and peripheral blood lymphocytes sensing conformational variants of prostate-specific antigen. J. Biol. Chem., 279, 51965-51972.
16. Ghahroudi, M.A., A. Desmyter, L. Wyns, R. Hamers, and S. Muyldermans. 1997. Selection and identification of single domain antibody fragments from camel heavy-chain antibodies. FEBS Letters, 414, 521-526.
17. Goujon, M., H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez. 2010. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res., 38 Suppl: W, 695-699.
* An introduction to the new European Bioinformatics Institute (EBI) online bioinformatics framework including integrated tools for sequence analysis. The site provides a common interface, and performs sequence input and parameter validation.
18. Li, W., A. Cowley, M. Uludag, T. Gur, H. McWilliam, S. Squizzato, Y.M. Park, N. Buso, and R. Lopez. 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res., 43(W1)W, 580-584.
* An update to the EBI platform that includes additional bioinformatics tools (e.g., RNA analysis, BLAST+) and database access (e.g., Pfam).
19. Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European molecular biology open software suite. Trends Genet., 16, 276-277.
* An introduction to the EMBOSS collection of integrated bioinformatics tools. Open-source licensing of EMBOSS permits free access so developers can build new sequence analysis tools using a variety of programming languages.
20. Sievers, F., A. Wilm, D.G. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J.D. Thompson, and D.G. Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol., 7, 539. doi:10.1038/msb.2011.75.
* An introduction to the Clustal Omega multiple sequence alignment tool. This tool was developed to align even extremely large numbers of sequences without sacrificing accuracy or execution time. Clustal Omega was here reported to be as accurate as the traditional alignment tools that perform well with small sequence numbers.
21. Wu, T.T. and E.A. Kabat. 1970. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J. Exp. Med., 132, 211-250.
22. Kabat, E.A., T.T. Wu, and H. Bilofsky. 1977. Unusual distribution of amino acids in complementarity-determing (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody combining sites. J. Biol. Chem., 252, 6609-6616.
23. Wesolowski, J., V. Alzogaray, J. Reyelt, M. Unger, K. Juarez, M. Urrutia, A. Cauerhff, W. Danquah, B. Rissiek, F. Scheuplein, N. Schwarz, S. Adriouch, O. Boyer, M. Seman, A. Licea, D.V. Serreze, F.A. Goldbaum, F. Haag, and F. Koch-Nolte. 2009. Single domain antibodies: Promising experimental and therapeutic tools in infection and immunity. Med. Microbiol. Immunol., 198, 157-174.