Ab-origin: An Improved Tool of Heavy Chain Rearrangement Analysis for Human Immunoglobulin

Xiaojing Wang[1,2], Wu Wei[1,2], SiYuan Zheng[2], Z.W. Cao[1,*], and Yixue Li[1,2,*]

[1] Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai, China
[2] Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences; Graduate School of the Chinese Academy of Sciences, 320 YueYang Road, Shanghai 200031, China
zwcao@scbit.org,
yxli@scbit.org

Abstract. An improved tool to explore the origin of human immunoglobulin heavy chains, named Ab-origin, has been developed. It can analyze in detail the V-D-J joints from the rearranged sequence by searching against germline databases. In addition to the known information about antibody recombination, appropriate score system and restriction of searching location are also incorporated to improve computing performance. When compared with a newly developed software SoDA, Ab-origin performed much better in both accuracy and stability, with 2, 7 and 1 percent higher for V, D and J respectively. Though only taking the human heavy chain for an example here, the algorithm also suits for the analysis of all immunoglobulin and TCR sequences.

Keywords: antibody diversity, V(D)J recombination.

1 Introduction

To protect ourselves from intruders, our immune system produces antibody proteins which are able to recognize and neutralize foreign substances, namely antigens. In response to the different varieties of antigens encountered over a human's lifetime, B cells need to make thousands of millions of different antibodies deriving from the limited immunologic information encoded in human genome[1]. It is estimated that the human body can produce at least 10^8 different antibodies[2], as a homo-dimmer of heavy and light peptide chains, each of them containing a unique variable region. In contrast to the huge diversities of unique regions of antibodies, the variable region of immunologic protein is only encoded by combination of three kinds of gene segments: variable (V), diversity (D) and joining (J) fragments (V and J segments only in the case of light chain). Taking heavy chain as an example, all the possible variable regions is only encoded by gene groups of 51 V genes, 27 D genes and 6 J genes in human chromosome 14 [2].

Obviously human beings has gained the most important and amazing immunological mechanism to generate the vast diversity of antibodies during the long

[*] Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part II, LNCS 4488, pp. 363–369, 2007.

history of evolution[1]. But how can the human bodies produce such a huge number of antibodies from limited number of gene groups? Several mechanisms in vivo have been revealed to answer this question, such as combinatorial V-(D)-J genetic joining, junctional flexibility, somatic hypermutation and combinatorial association of light and heavy chains[2]. It is notable that the random recombination of the V, (D), J genetic segments plays the most critical role[1], not only creating the diversity at level of 10^5 (10^3 for light chain), but also providing the basic structural frame for further diversity developing of the antibody variable region. In this sense, analysis of the V-(D)-J recombination process could help to facilitate antibody engineering for potential therapeutic and research applications.

With the accomplishment of human genome project, genetic information of immunoglobulins (Igs) has been collected into public databases[3], which makes bioinformatics analysis of V-D-J junction possible. Several tools have been developed trying to trace back to genetic coding from the mature immunoglobulin sequences, including some pioneer work which applied alignment methods[4, 5] or consecutive matches for D segment matching[6]. Recently, a new tool based on dynamic programming named SoDA was established, which is intended to process sequences in batch[7].

Although these existing tools produced some positive results, their methods were too complex. At the present time BLAST has been proved a very successful and efficient tool for sequence alignment, which motivates us to achieve the task with this handy and powerful tool. However, BLAST is a program for general sequence analysis, while the antibody sequences have their own specialities, so it is indispensable for us to assign these specialities to BLAST by setting appropriate parameters.

In this paper, a similar tool named Ab-origin was setup based on BLAST[8] algorithm, which has been widely accepted as a powerful tool for sequence alignment that allows custom parameter settings according to specific situations. Ab-origin was developed by JAVA language and run on Linux server. To better model the natural process of antibody maturation induced by antigen-affinity, the unconfirmed events such as D-D fusion[5, 9] and insertion/deletion during somatic hypermutations are excluded after checking with related reference [2, 10].

2 Method

2.1 Germline Database

Sequences of V, D, J germline genes of human immunoglobulin heavy chains were collected from IMGT[3] database. After removing the partial genes, the numbers of the V, D and J alleles are 187, 34, and 12 respectively. V germline sequences vary from 288 to 305 nucleotides in length, D vary from 11 to 37 and J from 48 to 63.

2.2 Principle

To our knowledge, V, D and J gene segments assemble through a site-specific recombination reaction which is thought to be a random assortment[11]. Selections of V, D and J segments during rearrangement process are independent. So, according to the Bayes' rule, there is

$$P(V,D,J|Q)=P(V)P(D)P(J)/P(Q)\,. \tag{1}$$

Where Q is the query sequence which is the target we need to analyze from mature antibody. V, D, and J represent the three germline segments, respectively. P(V,D,J|Q) is the probability of finding the correct V, D and J genes of the giving sequence Q. This is the real case from germline to the mature antibody sequence in our body, but when we decipher the mature sequence, the identification of the D region always lies on the location of the V and J segments which you found in the query sequence. So the formula changed to

$$P(V,D,J|Q)=P(V)P(D|V,J)P(J)/P(Q)\,. \tag{2}$$

When giving a observed sequence Q, P(Q) is a uniform. As a result, maximizing P(V), P(D|V,J), P(J) separately will also maximize the conditional probability P(V,D,J|Q). The probability of V, D and J was defined as a function of alignment score which we referenced from BLAST[12].

2.3 Algorithm

2.3.1 Search for V and J

Ab-origin calls NCBI BLAST software to find the best V gene segment in the library with the highest similarity to the query sequence. As we don't consider the insertion/deletion events, Ab-origin performs the alignment with a gap-forbidden style, moreover, uses a smaller word size and a specific scoring system of +5 for match and -6 for mismatch according to the similarity between germline and query sequence. For J segment the best hit is also found using the same method as the V segment. The probability was defined as the function of expected value getting from the BLAST process[12].

2.3.2 Search for D

As illustrated above, D germline sequences are shorter than the V, J germlines and vary largely in length. In the recombination, it was further shortened by deletions at both ends and heavily modified by somatic hypermutaion as the whole D was located in the CDR3 region. Because of its short length, false positive matches are of higher probability[12], and in some cases, no hit is found at all. For this reason, BLAST algorithm may not be effective enough to assess the D segment accurately; instead we try to use an algorithm extended from BLAST.

We defined the searching space as Q(V_end-5,J_start+5). The algorithm to find D germline in a query sequence is shown as following

```
For all D germlines do
 For every site n of query sequence from V_end-5 to
J_start+5 do
  For every site from n to n+D_length do
   If nucleotide of query sequence equals the one of
the germline sequence
      score=score+5;
   Else score=score-4;
   End if.
  End for.
 End for.
End for.
```

Then the scores are sorted to find the best one. To filter the stochastic matches, the hits with match number less than the half length of the D germline are discarded, and a rigorous penalty score -4 is adopted.

2.3.3 Identification of N and P Region

We searched the short palindromic sequence (P region) at the exact margin of the V, D and J region exactly reverse-complementary to the corresponding V, D and J germline, respectively. During the V(D)J joining process, a region of non-template nucleotides may be added by a terminal deoxynucleotidyl transferase (TdT) catalyzed reaction, namely N-region[2]. Ab-origin defined this region as the left parts of the query sequence after previous assignments.

2.4 Validation

To test Ab-origin, a simulation program was developed to generate artificial sequences of heavy chain variable regions. By randomly choosing V, D, J germline segments, the program simulates V-(D)-J rearrangement and to be more vivid, it cuts 0 to 5 nucleotides randomly from either sides of the joints of V-D and D-J combination due to the imprecise joining of the coding sequences [2]. Subsequently, up to 15 N-nucleotides were randomly chosen, adding to both the D-J and V-D joints. By introducing point mutations independently at each site, somatic hypermutation is simulated with a transition rate twice as much as a transversion rate[10]. Five different mutation rates were set ranging from 2% to 10% stepping 2% corresponding to different phase of antibody affinity maturation[2]. At each mutation rate, 1000 artificial sequences were generated. Finally, the sequences which contain termination codon (TAG, TGA, TAA) at current open reading frame (ORF) were removed.

3 Results and Discussions

Among the existing tools, only SoDA, IMGT/V-QUEST and VDJsolver can accept batch submission. SoDA has shown several advantages compared to other tools such as JOINSOLVER and V-QUEST, e.g. capacity of batch analysis, better results[7]. Thus we use simulated sequences to compare Ab-origin with SoDA for testing the performance of Ab-origin.

After filtration, numbers of the remaining simulation sequences without termination codons at five different mutation rates ranging from 2% to 10% stepping 2% are 353, 301, 243, 146, and 124 respectively. These sequences were analyzed by Ab-origin and SoDA using the same germline database from IMGT[3]. The results are shown in the Table1. At each mutation rate, Ab-origin identified more V, D, J segments correctly than SoDA did. The average accuracy ratios for V, D and J segments are up to 95%, 81% and 98%, respectively, compared to SoDA with 93%, 74% and 97%. Among which, the performance in determining D segment has the most notable improvment. In 353, 301, 243, 146 and 124 cases of analysis, Ab-origin correctly identified 292, 241, 193, 118 and 103 respectively, with an improvement of 7% in accuracy compared to SoDA.

Table 1. Results of five sets of simulated sequences with different mutation rates*

Mutation rate (%)		2		4		6		8		10	
Total number		353		301		243		146		124	
		SoDA	Ab[a]	SoDA	Ab	SoDA	Ab	SoDA	Ab	SoDA	Ab
Correct pickup	V	339	347	276	287	232	234	134	138	113	117
	D	276	292	235	241	178	193	111	118	84	103
	J	350	351	291	293	241	241	140	142	119	122

* The numbers represent the correct results from two programs at five different mutation rates. A correct inference means that the finding out gene segment was exactly the one we used in simulation, even not allowing for mismatched alleles. [a], Ab-origin.

In addition, we compared the error rates of the V, D and J segments at each mutation rate (Figure1). The figure shows a higher error rate in the D segments than that in V or J segments. The J segments were found to have the least error rates, which is due to limited choices in the antibody combinations and consistent with the previous studies[6, 7, 9]. Furthermore, all error rates of Ab-origin are lower than that of SoDA's at every mutation level (with an exception of the J segements at the mutation rate 6), and the variance of error rates among each mutation rate is also smaller (with standard deviation 0.016, 0.016 and 0.010 by Ab-origin and 0.023, 0.043 and 0.016 by SoDA for V, D and J segments, respectively). In summary, Ab-origin has better performance in both accuracy and stability, in contrast SoDA tends to have higher error rate concomitant with higher mutation rate which is also mentioned in their publication[7].

In the real case of natural antibody maturation process, the imprecise recombination of V-D and D-J segments may lead to the loss of several nucleotides in the D extremity, and meanwhile the random insertion of N-nucleotides may occur between the V-D and D-J joints. In addition, different D gene segments may have identical sequences, such as IGHD5-5*01 and IGHD5-18*01, IGHD4-4*01 and IGHD4-11*01, resulting in some unexpected inference mistakes. As a result, it's very difficult to recognize the D segment source. In contrast, the proportion of V and J being influenced in the recombination is much smaller, making the error rates in identifying D segment significantly higher than those of both V and J segments.

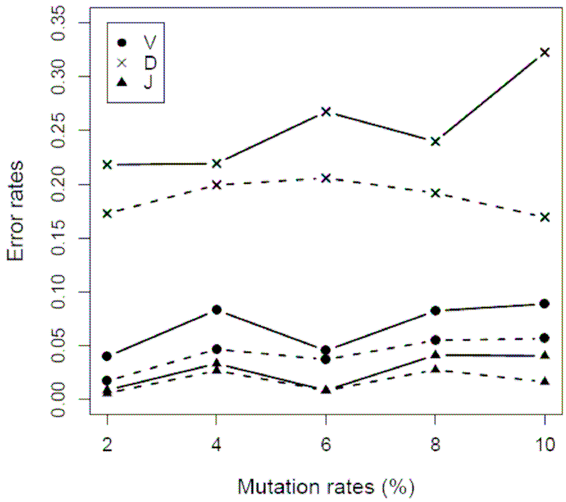

Fig. 1. Comparison of error rates between the results from Ab-origin (dashed line) and SoDA (solid line) at five mutation rates

In general, Ab-origin has better performance at every mutation level in both accuracy and stability; moreover, compared to SoDA it takes much shorter time to run (data not shown). Its accuracy and stability is due to effective parameter settings, while the fast running speed is duo to the optimization of BLAST in time and memory costs.

4 Conclusions

An improved tool Ab-origin was developed to efficiently identify V, D, and J gene segments from a rearranged antibody sequence by searching against germline database using appropriate rules. To evaluate the tool, we compared Ab-origin and SoDA with a set of artificial antibody sequences which were produced by simulating the antibody maturation process. The results show Ab-origin not only finds more correct V, D and J segments, with 2, 7 and 1 percent higher compared to SoDA respectively, but also reduces the computational cost. Though this paper only take the human heavy chain for an example, the algorithm established here also suits for the analysis of all immunoglobulin and TCR sequences in human, even other mammals which utilize similar antibody production mechanisms. Complementing the previous tools for partitioning the rearranged immunoglobulin sequences, Ab-origin may facilitate our understanding of antibody maturation process to provide the theoretical backgrounds for the antibody engineering for therapeutic and research applications.

Acknowledgments. This work was supported in part by grants from Ministry of Science and Technology China (2003CB715900, 2004CB720103), National Natural Science Foundation of China (30500107, 30670953), and Science and technology commission of Shanghai municipality (04DZ19850, 06PJ14072).

References

1. Maizels N.: Immunoglobulin gene diversification. Annu Rev Genet. Vol. 39. (2005) 23-46.
2. Goldsby R. A., Kindt T. J., Osborne B. A.Kuby J.: Chapter5, Immunology 5e. 5th edn; (2003).
3. Lefranc M. P.: IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. Vol. 29. (2001) 207-209.
4. Giudicelli V., Chaume D.Lefranc M. P.: IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res. Vol. 32. (2004) W435-440.
5. Corbett S. J., Tomlinson I. M., Sonnhammer E. L., Buck D.Winter G.: Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. J Mol Biol. Vol. 270. (1997) 587-597.
6. Souto-Carneiro M. M., Longo N. S., Russ D. E., Sun H. W.Lipsky P. E.: Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. J Immunol. Vol. 172. (2004) 6790-6802.
7. Volpe J. M., Cowell L. G.Kepler T. B.: SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. Bioinformatics. Vol. 22. (2006) 438-444.
8. Altschul S. F., Gish W., Miller W., Myers E. W.Lipman D. J.: Basic local alignment search tool. J Mol Biol. Vol. 215. (1990) 403-410.
9. Ohm-Laursen L., Nielsen M., Larsen S. R.Barington T.: No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. Immunology. Vol. 119. (2006) 265-277.
10. Odegard V. H.Schatz D. G.: Targeting of somatic hypermutation. Nat Rev Immunol. Vol. 6. (2006) 573-583.
11. Jung D., Giallourakis C., Mostoslavsky R.Alt F. W.: Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. Annu Rev Immunol. Vol. 24. (2006) 541-570.
12. Bedell J., Korf I.Yandell M.: BLAST. O'Reilly; (2003) 360.