ibc Interdisciplinary Bio Central

IBC 2011;3:14, 1-9 • DOI: 10.4051/ibc.2011.3.4.0014

Open Access

CONFERENCE PAPER

Global Sequence Homology Detection Using Word Conservation Probability

Jae-Seong Yang[1,‡], Dae-Kyum Kim[2,‡], Jinho Kim[2] and Sanguk Kim[1,2,3,*]

[1]School of Interdisciplinary Bioscience and Bioengineering, Pohang University of Science and Technology, Pohang, Korea
[2]Division of Molecular and Life Science, Pohang University of Science and Technology, Pohang, Korea
[3]Division of IT Convergence Engineering, Pohang University of Science and Technology, Pohang, Korea
[‡]These authors contributed equally to this work

Subject areas; Bioinformatics/Computational biology/Molecular modeling

Author contribution; Conceived and designed the experiments J-S.Y., D-K.K. and S.K.; Contributed reagents/ materials/ analysis tools J-S.Y., D-K.K. and J.K.; Wrote the paper J-S.Y., D-K.K. and S.K.

***Correspondence** and requests for materials should be addressed to S.K. (sukim@postech.ac.kr).

Editor; Keun Woo Lee, Gyeongsang National University, Korea

Received October 05, 2011
Accepted October 17, 2011
Published October 19, 2011

Citation; Yang, J-S., et al. Global Sequence Homology Detection Using Word Conservation Probability. IBC 2011, 3:14, 1-9. doi: 10.4051/ibc.2011.3.4.0014

Supporting online materials; Supporting online materials; http://www.ibc7.org/article/data_file.php?sid=270&mode=supplemental

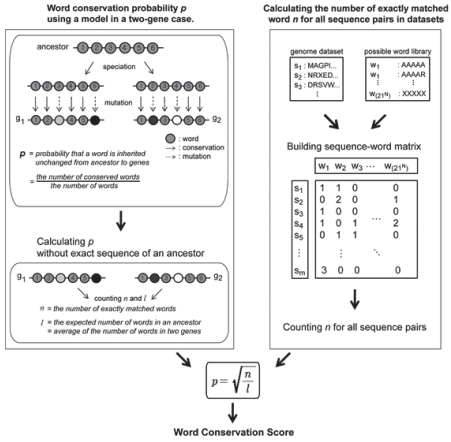
Funding; This work was supported by World Class University program (R31-2010-000-10100-0).

Competing interest; All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

© Yang, J-S. et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

SYNOPSIS

Protein homology detection is an important issue in comparative genomics. Because of the exponential growth of sequence databases, fast and efficient homology detection tools are urgently needed. Currently, for homology detection, sequence comparison methods using local alignment such as BLAST are generally used as they give a reasonable measure for sequence similarity. However, these methods have drawbacks in offering overall sequence similarity, especially in dealing with eukaryotic genomes that often contain many insertions and duplications on sequences. Also these methods do not provide the explicit models for speciation, thus it is difficult to interpret their similarity measure into homology detection. Here, we present a novel method based on Word Conservation Score (WCS) to address the current limitations of homology detection. Instead of counting each amino acid, we adopted the concept of 'Word' to compare sequences. WCS measures overall sequence similarity by comparing word contents, which is much faster than BLAST comparisons. Furthermore, evolutionary distance between homologous sequences could be measured by WCS. Therefore, we expect that sequence comparison with WCS is useful for the multiple-species-comparisons of large genomes. In the performance comparisons on protein structural classifications, our method showed a considerable improvement over BLAST. Our method found bigger micro-syntenic blocks which consist of orthologs with conserved gene order. By testing on various datasets, we showed that WCS gives faster and better overall similarity measure compared to BLAST.



Key Words: homology detection; alignment-free method; word conservation; global sequence homology; eukaryotic genome

BACKGROUND

Reliable homology detection is one of the central issues in comparative genomic analysis to assign the function annotations of proteins. Homologs are the genes that share common ancestry[1] which provide the clues of gene functions[2], detecting the homologous chromosomal region for comparative mapping[3], discriminating between coding regions from noncoding regions[4], and classifying protein family members[5].

There are two types of homology detection methods; alignment-based methods and alignment-free methods. Smith Waterman (SW) algorithm is one of the well known alignment-based methods to detect homologs with high sensitivity[6], however, it requires quadratic time for the sequence lengths. Thus, many efficient modifications of SW algorithm, such as FASTA[7] and BLAST[8], had been developed to save computation time[9,10]. BLAST[8] is probably the fastest and the most widely-used heuristic algorithm for sequence comparison and homology detection. However, BLAST is based on the local sequence alignment, which has inherent shortcomings. Local sequence alignment sacrifices investigation of global sequence homology for time reduction, which incurs problems in the comparisons of eukaryotic genes that have many gaps and insertions[11]. Eukaryotic genes often contain regions that have gene duplication and fusion, exon deletion and insertion[12-17]. A problem associated with local alignment-based homology detection is exemplified in Figure 1. Two homologous genes (g_1 and g_2) are originated from a common ancestor (dark gray), but have gone through different domain insertions in different sites of the genes (Figure 1A). The other gene (g_3, light gray) is not homologous with g_1 and g_2, but shares same domain insertion with g_2. If ancestry is properly considered, genes g_1 and g_2 should be detected as homologous sequences despite different domain insertions. But if local alignment was applied, g_2 and g_3 were falsely predicted as homologs, because later inserted domains are generally more conserved than the other part of gene that is inherited from ancestor and contain less gaps (Figure 1B). However, if overall similarity was properly considered, g_1 and g_2 were predicted as homologs (Figure 1C). Global sequence alignment between g_1 and g_2 can give three distinct short alignments. However, the three alignments were separate and have similarity scores calculated independently on each alignment in BLAST. To properly calculate overall similarity between two genes exemplified in the case of Figure 1C, a new sequence comparison method should be devised to reflect global alignments along the two sequences.

Moreover, local sequence alignment-based methods sacrifice overall sequence similarity, but they are not fast enough to find homologs from large sequence datasets such as eukaryotic genomes. For example, BLAST takes almost a day to find homologs between two eukaryote species for all-to-all comparisons. This impels researchers to develop a faster method to find homologs. Moreover, sequence comparison methods based on local alignment may not suitable for remote homology detection because they only provide weak relevance to protein function and structure[18,19]. Generally, if sequence identity is below twilight zones in sequence alignment (35%)[20], it often provokes unreliability of sequence alignment. Another drawback of alignment-dependent methods is that they do not provide rationale that explains why two sequences have a common ancestor, even though a proposed model is important to interpret similarity measure[21].

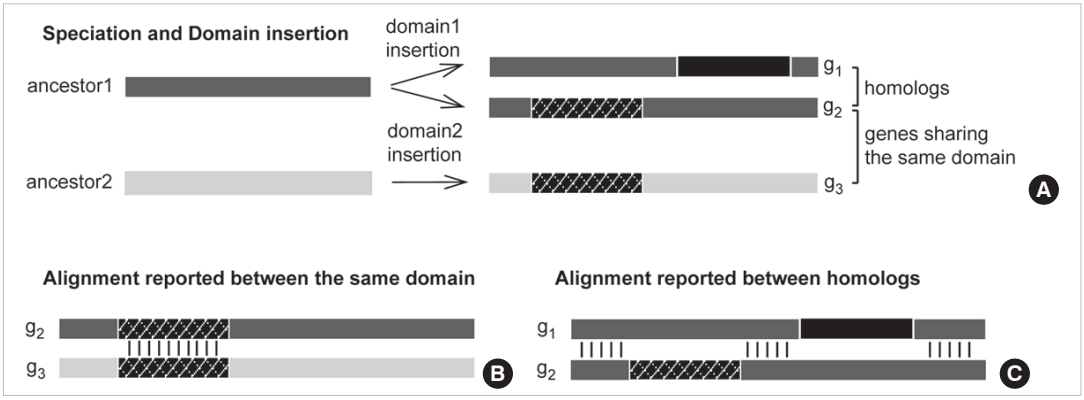


Figure 1. Problems in local alignment of sequences. (A) Model for speciation and domian insertion. g_1 and g_2 are genes speciated from ancestor(dark gray), but have different insertions (black box with no pattern and black box with checker pattern) on different sites. g_3 is not homologous to g_1 or g_2, but has same insertion (black box with checker pattern) as g_2. (B) Alignment between g_2 and g_3. In our model in panel A, since domains are inserted later than speciation, they are more-conserved and have longer aligned segment without gaps. So when we use local alignment, g_2 and g_3, not g_2 and g_1, can be predicted as homologs, which is not proper for homology detection. (C) Alignment between g_1 and g_2. When we use local alignment, we can detect one of three aligned segment, since insertions provoke gap penalty in extension step. However, all the aligned segments are considered in WCS score, because our method uses short exactly matched words to calculate overall sequence similarity.

Unlike alignment-based methods, alignment-free methods for homology detection use pre-built motif libraries or short sequence segments called words. It is widely known that motifs are located at functionally important regions of proteins such as catalytic site, binding sites, protein-protein interaction sites as well as structurally important sites[22,23]. Moreover, for predicting enzyme functions, programs based on a set of specific peptides and machine learning performed well[24] and fast[25] because they filtered redundant sequences. Although alignment-free methods have these advantages, but also have a number of shortcomings. Methods using motifs for homology detection required pre-built motif libraries. Also, machine learning-based methods needed a pre-defined training set for each homolog set.

Here, we present Word Conservation Score (WCS) in order to overcome the current limitations of sequence comparison methods. We assumed that if two sequences have a common ancestor then inherited words from a common ancestor are being mutated at the same rate in two sequences. With this assumption, our method measures their evolutionary distance from a common ancestor by comparing the word contents of the two sequences. By using the frequency of words in each dataset, *P*-values from WCS can give a confidence score that reflect the evolutional distance and the characteristics of two compared datasets. Moreover, since our method does not calculate sequence similarity through sequence alignment, but uses word contents in sequences, it dramatically reduced computation time. In addition, our method represents better overall similarity of sequences compared to local sequence comparison tools.

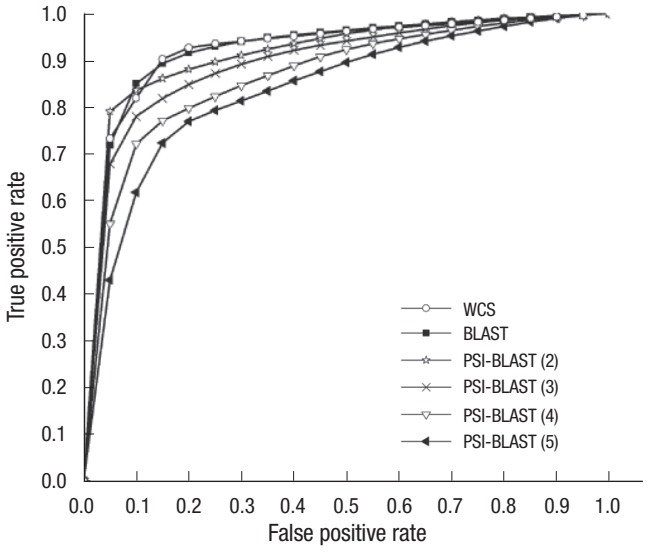


Figure 2. ROC curve of various methods based on SCOP annotation. We selected three methods BLAST, WCS and PSI-BLAST, and tested its performance based on SCOP annotation. All families were used for the performance test. The graph plots the ROC curve of each method. The number in a bracket next to PSI-BLAST indicates iteration time from PSI-BLAST.

RESULTS AND DISCUSSION

WCS reported more homologs in sequence comparison

We compared sequence comparison methods for all-to-all comparison of proteins that have SCOP annotation. Since WCS program does not require any training sets or information, methods that required training sets or multiple sequence alignment were not used for performance comparison. We used BLAST and PSI-BLAST for the test. In addition, we used all the SCOP families for the tests because a selection of families from SCOP could make bias on ROC scores.

We found that WCS program performed the fastest compared to other methods. Figure 2 and Table 1 summarized the performance of methods in ROC scores over all the SCOP families. WCS program, BLAST, and PSI-BLAST with 2 iterations showed similar ROC scores (0.93, 0.93, 0.92), which displayed that they have similar performance on protein structure classification. However, WCS program consumed 10-fold less computation time compared to other methods.

In addition, WCS showed better performance for remote homology detection, reporting 30 times more hits compared to other methods. As iteration continues, PSI-BLAST reported more hits but has lower ROC scores. In contrast, WCS showed up to ~6.4 times more hits and higher ROC scores (0.93 > 0.84) than those of PSI-BLAST on 5 iterations.

WCS showed significant time reduction for ortholog detection

To compare the running time of sequence comparison methods, we used whole protein sequence sets from KEGG/GENES database. BLAST has been widely used and often selected as a

Table 1. Result of remote homology detection on the SCOP benchmark database

Method	Hits	ROC score	Time (sec)	Time reduction (fold)
WCS	14,046,206	0.93	143.37	1
BLAST	544,611	0.93	2,420.58	16.88
PSI-BLAST(2)	1,043,506	0.92	4,906.77	34.22
PSI-BLAST(3)	1,516,997	0.9	8,093.91	56.45
PSI-BLAST(4)	1,907,552	0.87	11,703.14	81.63
PSI-BLAST(5)	2,211,365	0.84	15,802.20	110.22

The first column indicates the name of methods. Numbers in brackets on the right of PSI-BLAST mean iteration times for PSI-BLAST. The second column denotes the hits made from all-to-all comparison of SCOP database. Hits were defined as the sequence pairs which program reported similarity measure on a given option (see Method). The third column reports ROC score, the average area under the receiver operating curve (ROC). The fourth and fifth column report the computation time for all-to-all comparison of ASTRAL database and the ratio of computation time against that of WCS.

Table 2. Running time of BLAST-INPARANOID and WCS-INPARANOID, and their time reduction

Input species	Running time (sec)		Time reduction (fold)	Fraction of best matched ortholog groups in BLAST *F(BLAST, WCS)* (%)	Fraction of best matched ortholog groups in WCS *F(WCS, BLAST)* (%)
	BLAST-INPARANOID	WCS-INPARANOID			
Various bacteria	3,290.64 ± 45.19	210.90 ± 7.92	16.36 ± 0.52	98.01 ± 0.47	95.54 ± 1.24
Human, mouse	297,757.59	12,420.65	23.97	97.82	90.78
Human, dog	198,692.36	7,774.30	25.56	97.28	92.83
Human, *A. thaliana*	99,590.81	3,689.59	26.99	78.47	52.00
Human, worm	166,461.15	5,242.55	31.75	74.08	53.80
Human, worm, fly	145,873.66	2,276.43	64.08	73.76	58.52

The table shows computation times of each method for orthologs search and time reduction. Time reduction was obtained by dividing running time of BLAST-INPARANOID with one of WCS-INPARANOID. Last column presents fraction of best matched ortholog groups from BLAST-INPARANOID by WCS-INPARANOID, which means results from two methods are similar.

criterion to compare the computation time for tools reporting similarity measures[9,21]. Thus, we compared our method with BLAST.

As shown in Table 2, our method dramatically reduced computation time compared to BLAST-INPANARNOID. When we compared the two sequences of length l_1 and l_2, computation time for BLAST was proportional to the product of length of each sequence ($O(l_1 \times l_2)$). However, for WCS program, computation time of $O(l_1+l_2)$ was required to build a sequence-word matrix because there were $(l_1+l_2-2 \times N+2)$ words in two sequences, but constant time was requested to compare two sequences since there are always 21^N possible words. Thus, total time complexity of WCS program was $O(l_1+l_2)$, which was superior to that of BLAST. When we measured running time complexity, BLAST spent more than 1 day in most cases, however, WCS program took less than 4 hour and is ~60 times faster than BLAST. Also, as the size of target protein sequence sets grew bigger when we compared eukaryotic genomes, the time reduction became more significant.

WCS better reflected overall sequence similarity than BLAST

The ortholog detection results between remote species showed more significant difference in the members of ortholog groups (Table 2, fifth and sixth columns), when we compared the results of WCS-INPARANOID and BLAST-INPARANOID between human and worm, or between human and mouse in Table 2. It is certain that orthologs between remote species tend to be less similar than those between close species. Whereas, both methods showed similar performance in overall considering ROC scores based on SCOP annotation.

We further analysed the difference of the ortholog groups produced by WCS-INPARANOID and BLAST-INPARANOID. As a reference dataset, we used a manually curated dataset of orthologs that is obtained from the analysis of phylogenetic tree[26]. Ortholog groups from two methods were similar, in which 85.7% of curated groups and 74.1% of groups from BLAST-INPARANOID were matched to groups from WCS-INPARANOID. In addition, WCS-INPARANOID tended to report smaller size homolog groups and even some groups from BLAST-INPANRAOID split into many groups compared to WCS-INPARANOID.

Split groups reported by WCS-INPARANOID have higher overall sequence similarity and often contained homologs that had gaps in alignment which was described in Figure 1. For example, WCS-INPARANOID detected two homolog groups with more homolog members (13 > 8) of SOX family genes compared to BLAST-INPARANOID (Figure 3A). To assess overall similarity of each method, we made all-to-all sequence comparisons of the members within homolog groups. Both sequence identity and similarity within homolog groups for SOX family from WCS-INPARNOID were higher than BLAST-INPARANOID with statistical significance (using the t-test, *P*-value 6.20×10^{-6} and 1.45×10^{-9} for CLUSTALW identity and similarity, respectively) (Figure 3B). Also, we compared the ortholog alignments of SOX family genes from BLAST and WCS (Figure 3C). The alignment from BLAST-INPARANOID were found in small parts (55-130) of *SOXA_HUMAN* and no matches were found in C-terminal parts, although the missing C-terminal region of SOX gene was known to be essential for its function (Figure 3C)[27]. However, the alignment made from WCS-INPARANOID showed more significant matches along the sequence including C-terminal region, which suggested that WCS detected sequence regions with higher similarity.

Next, we confirmed that WCS detected overall sequence similarity better than BLAST algorithm. We compared the identity and similarity of sequence between homologs found from WCS and BLAST (Supplementary Figure 1). We found that correlation of sequence identity and WCS was stronger than that of BLAST ($R^2 = 0.70$ and $R^2 = 0.28$, respectively). We also observed the same tendency in sequence similarity.

Orthologs from WCS-INPARANOID showed better conservation of gene order than those from BLAST-INPARANOID

It has been shown that most orthologs tend to have conserved

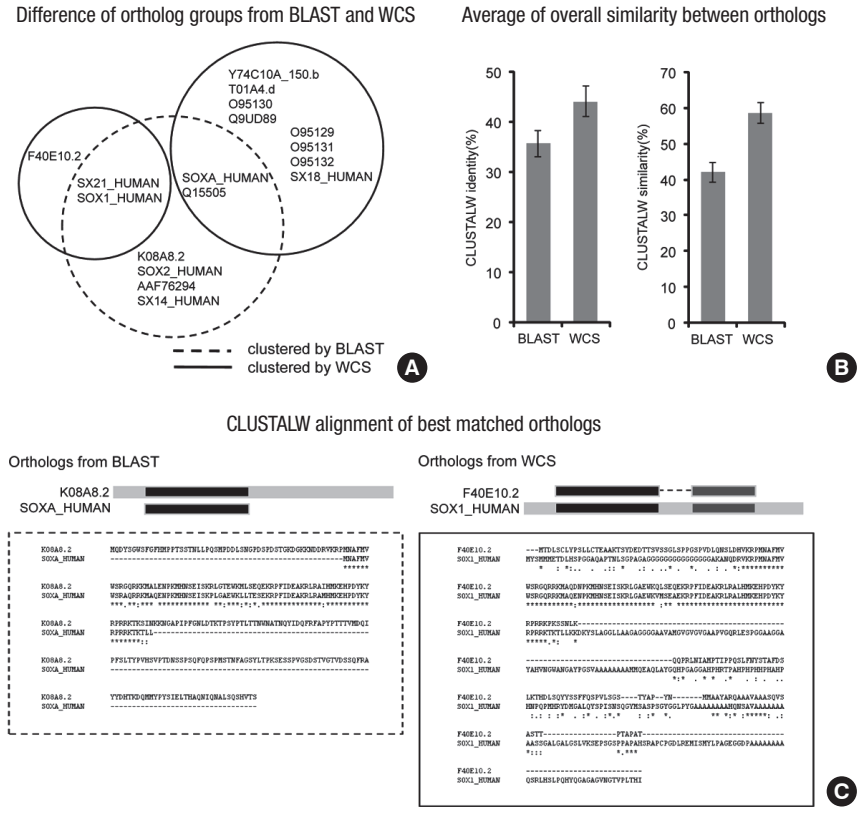


Figure 3. Comparison of ortholog alignment from SOX family from worm and human. (A) Comparison of SOX family ortholog groups from BLAST-INPARANOID and WCS-INPARANOID. (B) For each members of the ortholog group, we made all-to-all sequence alignment by using CLUSTALW. Averages of sequence identity and similarity in SOX family are presented, and both of them are significantly higher in sets from WCS-INPARANOID than BLAST-INPARANOID (*P*-value 6.20×10^{-6} and 1.45×10^{-9} for sequence identity and similarity, using the t-test). (C) Sequence alignment result of best matched orthologs in SOX family from both methods. Result in a dashed box is from BLAST-INPARANOID and the other in solid line is from WCS-INPARANOID. Alignment of BLAST result is more locally concentrated than that of WCS result as mentioned in Figure 1, and does not cover C-terminal region, which is known to be important for determining SOX's partner protein.

Table 3. Micro-Syntenic Block Sizes for BLAST-INPARNOID and WCS-INPARANOID

Species	Median of micro-syntenic block sizes		Number of genes in micro-syntenic block	
	BLAST-INPARANOID	WCS-INPARANOID	BLAST-INPARANOID	WCS-INPARANOID
Dog-human	5	7	16,549	17,243
Mouse-human	6	6	17,202	17,705
Yeast-human	2	2	90	234
Worm-human	2	2	606	687

gene order[28]. We compared the conservation of gene order as a benchmark for the reliability of ortholog detection. To measure the conservation of gene order, we compared the size of micro-syntenic blocks that have conserved gene order and transcriptional orientation[3]. We found that orthologs detected by WCS-INPARANOID have more conserved gene order than those of BLAST-INPARANOID within micro-syntenic blocks. Specially, the median of the size of micro-syntenic blocks from WCS-INPARANOID was bigger than that of BLAST-INPARANOID. The median of micro-syntenic block size and the total number of members in the blocks are reported in Table 3.

CONCLUSION

We found that WCS was faster in homology detection and detected better homologs in overall sequence similarity compared to local sequence alignment-based methods. Recently, genomes of multiple species are frequently used to find homologs. Because all-to-all comparison of sequences is a bottleneck for remote homology detection, we anticipate that WCS can significantly reduce computation time in sequence comparison for large dataset. Moreover, our method can find potential linear motifs as the frequently matched words in homolog groups. Therefore, we expect that these potential motifs from WCS could provide a clue for protein classification.

METHODS

Building sequence-word matrix and reducing amino acid alphabets

To reduce computation time, we built a sequence-word matrix, which contains the number of words in each sequence (the box on the right in Figure 4). If the word length is *N*, the total number of possible words in sequence is 21^N (20 amino acids and

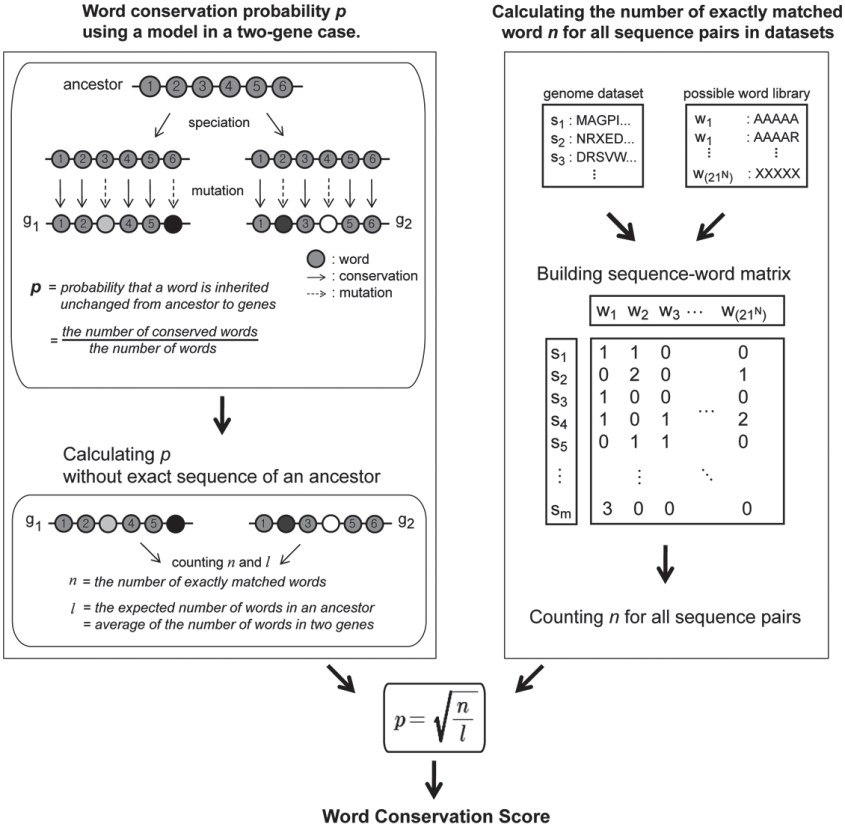


Figure 4. Overview of the method. In the box on the left, model of speciation is proposed to introduce word conservation probability. It is assumed that two sequences have a common ancestor and words are inherited unharmed from ancestor to sequences with the same word conservation probability *p*. However, we cannot construct a common ancestor perfectly from two sequences, statistical analysis is required to derive *p*. With statistical analysis, *p* is written by the number of exactly matched words *n* and the expected length *l* of a common ancestor, which is an average length of two sequences (see Method). In the box on the right, we built a sequence-word matrix for each database for fast comparison of two sequences. Scanning each sequence with a given word size *N*, we assigned (*i,j*)-*th* entry of matrix as the number of word (*j*) in the sequence (*i*). Using the matrix, we could count the number of exactly matched words to calculate the Word Conservation Score (WCS) rapidly for all sequence pairs in datasets.

Table 4. Reduced amino acid alphabets clustered by physicochemical properties

Size	Alphabet
20	I V L M T H Y F W N Q S D E A G K R P C
19	(IV) L M T H Y F W N Q S D E A G K R P C
18	(IV) (LM) T H Y F W N Q S D E A G K R P C
17	(IV) (LM) T H Y F W (NQ) S D E A G K R P C
16	(IV) (LM) T H Y F W (NQ) S (DE) A G K R P C
15	(IV) (LM) T H Y F W (NQ) S (DE) A G (KR) P C
14	(IV) (LM) T H Y F W (NQS) (DE) A G (KR) P C
13	(IV) (LM) T H Y F W (NQS) (DE) (AG) (KR) P C
12	(IV) (LM) T (HY) F W (NQS) (DE) (AG) (KR) P C
11	(IVLM) T (HY) F W (NQS) (DE) (AG) (KR) P C
10	(IVLM) T (HYF) W (NQS) (DE) (AG) (KR) P C
9	(IVLM) T (HYFW) (NQS) (DE) (AG) (KR) P C
8	(IVLM) T (HYFW) (NQSDE) (AG) (KR) P C
7	(IVLM) T (HYFW) (NQSDEAG) (KR) P C
6	(IVLMT) (HYFW) (NQSDEAG) (KR) P C

Clustered by Hierarchical Clustering Explorer with properties of amino acids.

other characters such as X). We collected all the words, moving word window by 1 amino acid, and constructed the vector for each sequence with 21^N dimensions, which represents the number of words for each kind. All the vectors are put together in a sequence-word matrix, which has (*i,j*)-*th* entry as the number of word (*j*) in the sequence (*i*).

Reducing amino acid alphabets is known to be effective to detect remote homology[29]. Thus, we constructed the property vector for each amino acid that includes physicochemical characteristics of amino acids, such as aliphaticity, aromaticity, charge, size, possibility of disulfide bond[30-33], pKa of N,C-terminal and side chain[33], non-polar surface area , estimated hydrophobic effect for residue burial and side chain burial[35], and pI[36]. With these property vectors, we clustered amino acids using Hierarchical Clustering Explorer[37]. The clusters of amino acids are shown in Table 4. Based on clustering options, we could group similar amino acids into same alphabet. For example, iso-leucine and valine are grouped as same alphabet.

Calculation of word conservation probability

To present an evolutionary distance from the speciation model, we defined word conservation probability *p*. Let us assume that two descendant genes g_1 and g_2 are diversified after speciation from a common ancestor, namely *ancestor* (the box on the left in Figure 4). We considered that *ancestor* would have *l* words which is average words counts of g_1, and g_2 (l_1 and l_2, respectively). From *ancestor*, the specific portion of words would be conserved and handed down to g_1 and g_2, and let this probability that each word would not change through evolution as the word conservation probability *p*. As the evolutionary distance

increases, less words are conserved and p become smaller.

Although prediction of precise common ancestor was impossible, we could provide a relevant evolutionary distance between g_1 and g_2 by presenting p with the number of exactly matched words n. Assumed that all the words with exact match between g_1 and g_2 originated from their common ancestor ($n \leq l$), the number of conserved words between common ancestor and g_1 can be estimated as $l \times p$. In turn, we estimated the number of the same words between g_1 and g_2 as $l \times p^2$. Using this possibility, we presented p as

$$p = \sqrt{\frac{n}{l}}$$

For a practical purpose, we defined WCS as $-1{,}000 \times \log_{10}(1-p)$. Thus, the bigger WCS indicated the closer evolutionary distance in this model.

Calculation of statistical significance for WCS

To provide a statistical significance of WCS, we evaluated *P*-value as the probability of obtaining a sequence that has same or more exactly matched words than the compared sequence pair[38]. *P*-values of WCS depend on not only two compared sequences, but also other sequences in the dataset. When the sequence of g_1 is a query sequence in dataset DS_1 and the sequence of g_2 in dataset DS_2, we could estimate the probability λ that any words in sequence of g_1 were randomly matched in dataset DS_2.

$$\lambda = \frac{\text{the number of } g_1 \text{ words in } DS_2}{\text{the number of words in } DS_2}$$

Depending on the word contents of sequences in DS_2, λ can be changed.

We used Poisson distribution to find probability $f(i; \lambda l)$, that a random sequence has i exactly matched words for given λ and l. That is,

$$f(i;\lambda l) = \frac{(\lambda l)^i e^{-\lambda l}}{i!}$$

Thus *P*-value for n exactly matched words of g_1 and g_2 is

$$P = 1 - \sum_{i=0}^{n-1} f(i;\lambda l)$$

To reduce the computation time of factorial in $f(i; \lambda l)$, we used sterling's approximation[39], which is

$$k! \approx e^{-k} k^k \sqrt{2\pi k}$$

Homolog clustering scheme

We clustered homologous sequences between *Homo sapiens* and *Caenorhabditis elegans* by using WCS-INPARANOID, which is a modified INPARANOID[40] with WCS instead of BLAST bit score as a sequence similarity measure. We compared our result with INPARANOID with BLAST bit score (BLAST-INPARANOID) on a manually curated dataset of orthologs which is obtained by the analysis of phylogenetic trees[26]. We found the best matched set from WCS-INPARANOID with the manually curated dataset. We also found the largest number of common members from WCS-INPARANOID compared with the manually curated dataset. To define the performance, we introduced the fraction of best matched homolog groups generated by method *X* and *Y*, *F(X,Y)* (Supplementary Figure 2A). Performance was defined as multiplication of the fraction of the best matched groups from WCS-INPARANOID and BLAST-INPARANOID to object function for performance was *F(BLAST, WCS)* × *F(manual curation, WCS)* (Supplementary Figure 2B). The best performance based on the human and worm genomes drawn when the option was adjusted as group number 17 and score cut-off 40 (the darkest area of Supplementary Figure 2B).

Datasets

We applied our method to the whole protein coding gene set of the species from KEGG/GENES database[41,42] to test the efficiency of our method (Database used for the test can be downloaded from the program homepage). To compare the accuracy of homology detection, we used the whole genomes of *Homo sapiens* and *Caenorhabditis elegans* that were used as a test set of INPARANOID[26,40]. To obtain ROC scores, protein sequence sets with less than 95% sequence identity of protein database and SCOP annotation from version 1.60 were used.[43,44] Positions of genes on chromosomes are obtained from Ensembl to analyze orders and micro-syntenic block.

Parameters and test environment

The performance of our method is compared with the following methods: BLAST-INPARNOID, PSI-BLAST, BLAST, and CLUSTALW. None of them required any previously aligned Multiple Sequence Alignment (MSA) files or training sets. The setup procedures are described as follows.

For BLAST-INPARANOID, databases with raw sequences in FASTA format were used as an input. We used default parameters of BLAST to obtain bit score in INPARANOID (scoring matrix = BLOSUM62, e = 10.0), and the default parameters for INPARANOID (score cut-off = 20, sequence overlap cut-off = 0.5).

For BLAST and PSI-BLAST, we used the same strategy (scoring matrix = BLOSUM62, e = 10.0). For PSI-BLAST, at each round, we made differences in the iterations from 2 to 5. Also for CLUSTALW, we used version 2.0 with default options.

To measure computation time for each method, we used a computer with Intel 2,660 MHz CPU and 30 GB memory. We

implemented WCS program based on the source code from an open source project, Cd-hit[25]. All the performance tests were conducted with word length 5 which is default option of WCS program.

ACKNOWLEDGEMENTS

This work was supported by World Class University program (R31-2010-000-10100-0). We thank the SBI members for critical discussions and careful review of this manuscript.

REFERENCES

1. Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool* 19, 99-113.
2. Wu, C.H., Huang, H, Yeh, L.S., and Barker, W.C. (2003). Protein family classification and functional annotation. *Comput Biol Chem* 27, 37-47.
3. Goodstadt, L., and Ponting, C.P. (2006). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol* 2, e133.
4. Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., and Lander, E.S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* 104, 19428-19433.
5. Redfern, O., Grant, A., Maibaum, M., and Orengo, C. (2005). Survey of current protein family databases and their application in comparative, structural and functional genomics. *J Chromatogr B Analyt Technol Biomed Life Sci* 815, 97-107.
6. Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. *J Mol Biol* 147, 195-197.
7. Lipman, D.J., and Pearson, W.R. (1985). Rapid and sensitive protein similarity searches. *Science* 227, 1435-1441.
8. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403-410.
9. Itoh, M., Goto, S., Akutsu, T., and Kanehisa, M. (2005). Fast and accurate database homology search using upper bounds of local alignment scores. *Bioinformatics* 21, 912-921.
10. Moreno-Hagelsieb, G., and Latimer, K. (2008). Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24, 319-324.
11. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (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, 4673-4680.
12. Jones, C.D., Custer, A.W., and Begun, D.J. (2005). Origin and evolution of a chimeric fusion gene in *Drosophila subobscura, D. madeirensis* and *D. guanche*. *Genetics* 170, 207-219.
13. Sayah, D.M., Sokolskaja, E., Berthoux, L., and Luban, J. (2004). Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* 430, 569-573.
14. Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat Rev Genet* 4, 865-875.
15. Fumasoni, I., Meani, N., Rambaldi, D., Scafetta, G., Alcalay, M., and Ciccarelli, F.D. (2007). Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates. *BMC Evol Biol* 7, 187.
16. Ben-Shlomo, I., Yu Hsu, S., Rauch, R., Kowalski, H.W., and Hsueh, A.J. (2003). Signaling receptome: a genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. *Sci STKE* 2003: RE9.
17. Alexeyenko, A., Tamas, I., Liu, G., and Sonnhammer, E.L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* 22, e9-15.
18. Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* 333, 863-882.
19. Hegyi, H., and Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* 288, 147-164.
20. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng* 12, 85-94.
21. Hochreiter, S., Heusel, M., and Obermayer, K. (2007). Fast model-based protein homology detection without alignment. *Bioinformatics* 23, 1728-1736.
22. Ben-Hur, A., and Brutlag, D. (2003). Remote homology detection: a motif based approach. *Bioinformatics* 19 Suppl 1, i26-33.
23. Tong, A.H., Drees, B., Nardelli, G., Bader, G.D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., et al. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295, 321-324.
24. Kunik, V., Meroz, Y., Solan, Z., Sandbank, B., Weingart, U., Ruppin, E., and Horn, D. (2007). Functional representation of enzymes by specific peptides. *PLoS Comput Biol* 3, e167.
25. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658-1659.
26. Remm, M., and Sonnhammer, E. (2000). Classification of transmembrane protein families in the *Caenorhabditis elegans* genome and identification of human orthologs. *Genome Res* 10, 1679-1689.
27. Kamachi, Y., Cheah, K.S., and Kondoh, H. (1999). Mechanism of regulatory target selection by the SOX high-mobility-group domain proteins as revealed by comparison of SOX1/2/3 and SOX9. *Mol Cell Biol* 19, 107-120.
28. Hurst, L.D., Pal, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* 5, 299-310.
29. Ogul, H., and Mumcuoglu, E.U. (2007). A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets. *Biosystems* 87, 75-81.
30. Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature* 277, 491-492.
31. Wolfenden, R., Andersson, L., Cullis, P.M., and Southgate, C.C. (1981). Affinities of amino acid side chains for solvent water. *Biochemistry* 20, 849-855.
32. Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157, 105-132.
33. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H.

(1985). Hydrophobicity of amino acid residues in globular proteins. *Science* 229, 834-838.
34. Massey, K.A., Blakeslee, C.H., and Pitkow, H.S. (1998). A review of physiological and metabolic effects of essential amino acids. *Amino Acids* 14, 271-300.
35. Karplus, P.A. (1997). Hydrophobicity regained. *Protein Sci* 6, 1302-1307.
36. Windholz, M. (1984). The Merck Index Online. *Science* 226, 1250.
37. Seo, J., Gordish-Dressman, H., and Hoffman, E.P. (2006). An interactive power analysis tool for microarray hypothesis testing and generation. *Bioinformatics* 22, 808-814.
38. Edwards, A.W. (1969). Statistical methods in scientific inference. *Nature* 222, 1233-1237.
39. Whittaker, E.T., and Robinson, G.(1967). The calculus of observations; an introduction to numerical analysis, 4th edition., (New York: Dover Publications).
40. Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314, 1041-1052.
41. Kanehisa, M. (2002). The KEGG database. *Novartis Found Symp* 247, 91-101; discussion 101-103, 119-128, 244-152.
42. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32, D277-280.
43. Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536-540.
44. Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 36, D419-425.