February 2010

Dissertation for Doctor of Science

Probabilistic Models for Inferring Orthology in Comparative Genome Analysis

Keio University
Graduate School of Science and Technology
School of Fundamental Science and Technology
Center for Biosciences and Informatics

Tsuyoshi Hachiya

Contents

Abbreviation

AGG,	asymmetric generalized Gaussian
DAG,	directed acyclic graph
DPD,	diagonal pseudo distance
FoSTeS,	fork stalling template switching
GG,	generalized Gaussian
KAAS,	KEGG Automatic Annotation Server
KO,	KEGG Orthology
Mbo,	*M. bovis*
Mle,	*M. leprae*
MMEJ,	microhomology-mediated end-joining
MMIR,	microhomology/microsatellite-induced replication
Mpa,	*M. avium*
Mtu,	*M. tuberculosis*
NAHR,	nonallelic homologous recombination
OG,	orthologous gene
OGG,	one-sided generalized Gaussian
ORF,	open reading frame
PDF,	probability density function
RBH,	reciprocal best BLAST hit
RBS,	ribosomal binding site
RSD,	reciprocal smallest distance
WNNSO,	weighted number of neighboring seed orthologs

Chapter I
Introduction

Dramatic increases in the throughput of nucleotide sequencing machines yield petabyte-scale data sets of biological sequences (Cochrane *et al.*, 2009). The number of gene sequences is being exponentially increased as well as the number of sequenced genomes (Koonin and Wolf, 2008). As of this writing (November 2009), 1,015 prokaryotic genomes and 121 eukaryotic genomes have been sequenced according to the GOLD database (Liolios *et al.*, 2008). Current collection of sequenced genomes provides us with new opportunities and challenges in comparative genomics and evolutionary biology. Among the new challenges, we chose to focus our research on the following theme: inference of evolutionary relationship among biological sequences. In order to mine valuable insights from the rapidly growing repository of biological sequences, determining evolutionary relationships is one of the most fundamental prerequisites (Alexeyenko *et al.*, 2006; Hulsen *et al.*, 2006; Chen *et al.*, 2007).

Evolutionary relationships among biological sequences are divided into two classes: *orthology* and *paralogy* (Fitch, 1970). A certain genomic region of a species begins to take different evolutionary paths when the species is speciated or the genomic region is duplicated. Orthology refers to evolutionary relationship between biological sequences evolved by speciation, whilst paralogy refers to evolutionary relationship between biological sequences evolved by duplication.

In this dissertation, we propose two statistical algorithms to infer orthologous relationships among biological sequences based upon *probabilistic models* and a theory named *decision theory*.

The first algorithm was designed to accurately infer orthologous relationship of chromosomal segments among different genomes (Hachiya *et al.*, 2009). Exponential growth of sequenced genomes makes it possible to study chromosome-level mutations such as genome rearrangements (Pevzner and Tesler, 2003a) and segmental duplications (Jiang *et al.*, 2008) as well as nucleotide-level mutations such as insertions, deletions, and substitutions. Accurate identification of orthologous chromosomal segments among different genomes is essential for the analyses of chromosome-level mutations (Bourque *et al.*, 2004, 2005).

The second algorithm was designed to reveal the relationship between the gene order along chromosomes and the biological functions of the genes (Hachiya and Sakakibara, 2009). It was revealed that, in mammals, 39 Hox genes are clustered on four chromosomal loci, and their order along the chromosomes is correlated with their spatial pattern of expression along the anterior-posterior and proximal-distal axes (Chang, 2009). In prokaryotes, the proteins encoded by neighboring genes along the chromosome are likely to physically interact with each other (Dandekar *et al.*, 1998). The rapid increase of the availability of sequenced genomes provides us with an opportunity to explore the relationship between chromosomal position and biological

function of genes by analyzing in a systematic and genome-scale manner rather than by focusing on a few genes.

1 Inference, Probabilistic Models and Decision Theory

Inference is the process of drawing a conclusion by applying rules or theories to observations or hypotheses (MacKay, 2003). Statistical methodologies in bioinformatics to make inferences have been applied to the observations obtained from biological experiments. For example, gene prediction algorithms have been applied to genomic sequences observed by using sequencing machines (Lowe and Eddy, 1999; Delcher *et al.*, 2007; Baten *et al.*, 2008; Wang and Ruvinsky, 2009). It is noted that if a conclusion drawn by an inference process can be easily observed from biological experiments, the inference process does not play an important role in biology. Thus, inference processes exert more beneficial impact on biology in the case where the conclusions drawn by the inference processes are intrinsically not observable, or difficult to be observed. In evolutionary biology, there are many concepts that are inherently not observable: evolutionary histories in the past, purifying selections to maintain protein sequences, and the fitness of an organism to an environment (Pigliucci and Kaplan, 2006). For this reason, statistical methodologies to make inferences play a key role in evolutionary biology (Durbin *et al.*, 1998; Nei and Kumar, 2000; Yang, 2006).

In this dissertation, we employ a statistical framework based upon probabilistic models and the decision theory to make inferences. Probabilistic models describe the probability density of a random variable of interest, or the joint probability density of a set of random variables (Durbin *et al.*, 1998; Bishop, 2006). In the case of gene-finding algorithms, for example, joint probability densities of the occurrence of a certain k-mer nucleotides in protein-coding regions and in non-coding regions are respectively described by some probabilistic models such as Markov chain models and interpolated Markov models (Delcher *et al.*, 2007).

Let s_i be the DNA sequence of the i^{th} open reading frame (ORF) candidate, $P(s_i|\text{coding})$ be the occurrence probability of s_i in protein-coding regions, and $P(s_i|\text{non-coding})$ be that in non-coding regions. Assume that $P(s_i|\text{coding})$ and $P(s_i|\text{non-coding})$ are given for each possible DNA sequence s_i. Then, the decision theory can be applied to the discrimination between protein-coding regions and non-coding regions. For example, the logarithm of the ratio of the two occurrence probabilities is useful to score the sequence of the i^{th} ORF candidate:

$$score(s_i) = \log \frac{P(s_i|\text{coding})}{P(s_i|\text{non-coding})} \ . \tag{I.1}$$

When the base of logarithm is 2, the score given by Eq. (I.1) is in bit-scale. The decision theory states that the following rule minimizes the expected number of misclassifications estimated based on the two occurrence probability densities (Bishop, 2006):

- If $score(s_i) \geq 0$, then the i^{th} ORF candidate is a coding region.
- Otherwise, the i^{th} ORF candidate is a non-coding region.

In addition, this framework enables to make inferences based on heterogeneous information. For example, in the case of gene-finding algorithms, not only the sequence of ORF candidates but also the sequence of ribosomal biding site (RBS) candidate are useful to distinguish protein-coding regions from non-coding regions (Delcher *et al.*, 2007). Let r_i be the sequence of the RBS candidate located on the up-stream of the i^{th} ORF candidate, $P(r_i|\text{RBS})$ be the occurrence probability of r_i in RBS regions, and $P(r_i|\text{non-RBH})$ be that in non-RBS regions. Then, the score of sequence of an ORF candidate and the score of the sequence of the RBS candidate can be simply added as shown in Eq. (I.2) because both scores are in bit-scale.

$$\begin{aligned} score(i^{\text{th}}\ \text{ORF}) &= score(s_i) + score(r_i) \\ &= \log \frac{P(s_i|\text{coding})}{P(s_i|\text{non-coding})} + \log \frac{P(r_i|\text{RBS})}{P(r_i|\text{non-RBS})} \ . \end{aligned} \tag{I.2}$$

In order to minimize the expected number of misclassifications, the decision theory gives rise to the following rule:

- If $score(i^{\text{th}}\ \text{ORF}) \geq 0$, then the i^{th} ORF candidate is a coding region.
- Otherwise, the i^{th} ORF candidate is a non-coding region.

2 Probabilistic Models for Inferring Orthologous Segments

When comparing gene orders of two closely related genomes, G_A and G_B, genomic segments, in which the gene order of G_A is the same as that of G_B, can be found. Let G_0 denote an ancestral genome of G_A and G_B. From the parsimonious viewpoint, it is elucidated that the gene order of the corresponding genomic regions on G_0 would be identical to those of G_A and G_B, and that any genome rearrangement (e.g. inversion, fusion, fission and translocation), gene duplication, and gene insertion/deletion would not change the gene order of the genomic regions during the evolutionary histories from G_0 to G_A and from G_0 to G_B.

The term *orthologous segment* is defined as a set of genomic segments in different organisms descended from a common ancestor without any rearrangements (Dewey *et al.*, 2006): the genomic segments in which the gene order on the ancestral genome has not been disrupted by any genome rearrangement during the evolutionary histories from the ancestral genome to the descendant genomes under comparison. It is noted that the evolutionary histories from an ancestral genome to descendant genomes are intrinsically not observable. For this reason, in order to detect orthologous segments, an inference process is required.

Accurate detection of orthologous segments among multiple genomes is essential for the following subsequent analyses: inferring rearrangement-based phylogenies (Tesler, 2002; Bourque *et al.*, 2004), reconstructing ancestral genomes (Bourque *et al.*, 2005;

Murphy *et al.*, 2005; Ma *et al.*, 2006), computing whole genome alignments (Dewey *et al.*, 2006; Gibbs *et al.*, 2004; Waterston *et al.*, 2002), identifying orthologous genes (Hubbard *et al.*, 2005; Zheng *et al.*, 2005), and detecting non-coding functional elements such as regulatory elements (Frazer *et al.*, 2004). Thus, a number of algorithms to detect orthologous segments have been proposed (Vandepoele *et al.*, 2002; Calabrese *et al.*, 2003; Cannon *et al.*, 2003; Kent *et al.*, 2003; Pevzner and Tesler, 2003a; Haas *et al.*, 2004; Soderlund *et al.*, 2006; Sinha and Meller, 2007). A general strategy that is common among almost the current algorithms is as follows.

Step 1 Take multiple genome sequences, $G_1, G_2, \ldots, G_n (n \geq 2)$, as input.

Step 2 Detect short genomic regions with high similarity among all genomes, $G_1, G_2, \ldots, G_n$. These genomic regions are referred to as *anchors*.

Step 3 Detect genomic regions in which the order of anchors is well conserved among all genomes, $G_1, G_2, \ldots, G_n$.

Step 4 If the genomic regions detected in the Step 3 contain anchors more densely than a certain threshold of anchor density, output the genomic segments as orthologous segments.

A certain fraction of anchors are mapped on non-orthologous genomic regions because repeat and paralogous sequences also generate anchors as well as orthologous sequences do. In order to detect orthologous segments while filtering out non-orthologous anchors, the above strategy implements an inference process based upon the following two hypotheses.

The first hypothesis is the *parsimonious hypothesis*. This hypothesis assumes that if the order of anchors in a certain genomic region is well conserved among all genomes, $G_1, G_2, \ldots, G_n$, the order of corresponding anchor sequences on the ancestral genome is the same as the descendant genomes, and that any genome rearrangement has not disrupted the order of anchors during evolutionary histories from the ancestral genome to the descendant genomes under comparison. The second hypothesis is the *density hypothesis*. The density hypothesis assumes that anchors are more densely mapped on orthologous segments than on the other genomic regions. Based upon these two hypotheses, the above strategy distinguishes orthologous segments from the other genomic regions.

A drawback of all existing algorithms is that they do not equip with a framework to computationally determine an appropriate threshold of the anchor density to distinguish orthologous segments from the other genomic regions; the threshold is required to be given by users. In order to make accurate inference of orthologous segments, it is needed to model two respective anchor densities for orthologous segments and for other genomic regions, and to optimize the threshold of anchor density by making use of the two models based on the decision theory. Another drawback of almost existing programs is that they are implemented to compare pairwise genomes; they can not be applied to the identification of orthologous segments among multiple genomes.

In Chapter II, we describe a novel algorithm named OSfinder (Orthologous Segment finder) (Hachiya *et al.*, 2009). OSfinder infers orthologous segments among multiple genomes by modeling the respective anchor densities for orthologous segments and for

other genomic regions based on probabilistic models, and determining the threshold of anchor density based on the decision theory. Thus, OSfinder makes it possible to automatically optimize the threshold of anchor density for each set of genomes. This automation would improve the throughput of comparative genomic analyses because manual optimization of the threshold value requires the computation of orthologous segments multiple times while varying the threshold values. Moreover, our evaluation tests using mammalian and bacterial genomes demonstrated that OSfinder shows higher accuracy than existing algorithms. This result implies that it is difficult to achieve a high accuracy based on manually-defined threshold of anchor density, and that the use of probabilistic models and decision theory improves the accuracy of the identification of orthologous segments. Furthermore, the accuracy of OSfinder in multiple genome comparisons is greater than that in pairwise genome comparisons. This result suggests that the accuracy of identifying orthologous segments would be increased as the number of sequenced genomes increases.

3 Probabilistic Models for Inferring Positional Orthologs

It is widely accepted that orthologous genes (also referred to as *orthologs*) have identical function to each other (Ohno, 1970; Remm *et al.*, 2001; Chen *et al.*, 2006), whereas paralogous genes (also referred to as *paralogs*) have different biological functions (Ohno, 1970; Zhang *et al.*, 1998; Moore and Purugganan, 2003; Rodriguez-Trelles *et al.*, 2003; Thornton and Long, 2005; Han *et al.*, 2009). Accordingly, identifying orthologs and paralogs is an effective way to predict gene functions and to understand the radiation of gene families, respectively (Remm *et al.*, 2001; Li *et al.*, 2003; Hulsen *et al.*, 2006).

Positional orthologs are referred to as genes in different species that are orthologous to each other and are located on corresponding chromosomal positions. Suppose that two genomes, G_A and G_B, are speciated from a common ancestor G_0, and the gene order of three neighboring genes have not been disrupted during evolutionary histories from G_0 to G_A and from G_0 to G_B. Let the descendant of the neighboring gene cluster in G_A and G_B be $\{a_{i-1}, a_i, a_{i+1}\}$ and $\{b_{i-1}, b_i, b_{i+1}\}$, respectively. In addition, suppose that b_i was duplicated after the speciation of G_A and G_B, and G_B comes to encode a new gene b'_i as shown in Fig. I.1. In this case, there are three positional orthologs: (a_{i-1}, b_{i-1}), (a_i, b_i), and (a_{i+1}, b_{i+1}). Although the gene pair (a_i, b'_i) is an ortholog, it is not a positional ortholog because the two genes are not located on corresponding chromosomal positions.

Although a number of algorithms have been proposed to identify orthologs (Tatusov *et al.*, 1997; Remm *et al.*, 2001; Li *et al.*, 2003; Tatusov *et al.*, 2003; Dehal and Boore, 2006; Vilella *et al.*, 2009), there are a few algorithms to identify positional orthologs (Fu *et al.*, 2007; Rödelsperger and Dieterich, 2008). However, recent researches on the relationship between gene order and gene function in prokaryotic genomes reveal that positional orthologs tend to have identical function to each other, and the other types of orthologs (e.g. in-paralogs) tend to have different biological

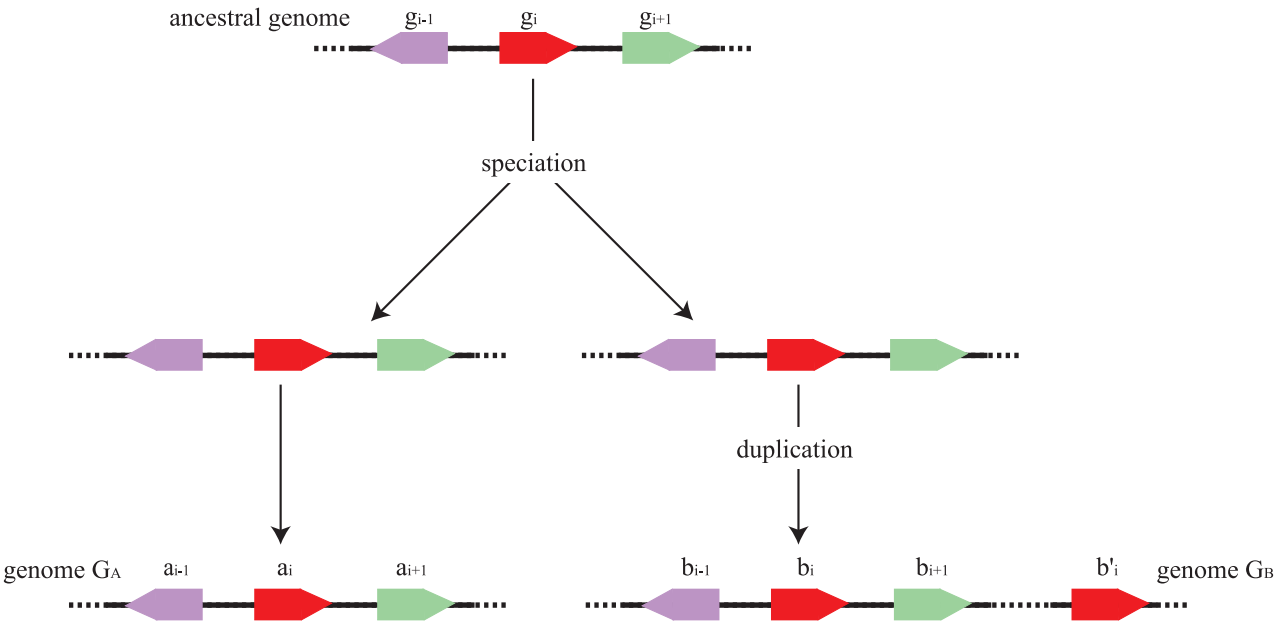


Fig. I.1 An illustration of a genome evolution with a duplication event. We here suppose that two genomes, G_A and G_B, are speciated from a common ancestor, and the gene order of three neighboring genes have not been disrupted. The descendant of the gene cluster in G_A and G_B are denoted as $\{a_{i-1}, a_i, a_{i+1}\}$ and $\{b_{i-1}, b_i, b_{i+1}\}$, respectively. In addition, we suppose that b_i is duplicated after the speciation of G_A and G_B, and G_B comes to encode a new gene b'_i.

functions (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999a,b; Snel *et al.*, 2000; Notebaart *et al.*, 2005). Furthermore, the identification of positional orthologs is useful to detect *conserved gene clusters*. A conserved gene cluster is defined as a cluster of neighboring genes whose gene order is conserved across several species. Since the proteins encoded by the genes in conserved gene clusters have shown to tend to physically interact with each other (Dandekar *et al.*, 1998), the detection of conserved gene clusters provides valuable information to predict protein-protein interactions (Huynen *et al.*, 2000; Wolf *et al.*, 2001; Li *et al.*, 2007). It is noted that the identification of positional orthologs is useful to predict molecular functions of a protein, whereas the detection of conserved gene clusters is useful to predict a higher order function of genes (e.g. with which other protein it interacts) (Huynen *et al.*, 2000).

In order to detect positional orthologs, it is needed to develop an algorithm which takes into account not only the conservation of protein sequences but also gene order conservation from the definition of positional orthologs. Existing algorithms to identify positional orthologs (Fu *et al.*, 2007; Rödelsperger and Dieterich, 2008) begin with detecting homologous gene pairs (e.g. gene pairs whose bit score is greater than 50 bits in BLASTP comparison (Altschul *et al.*, 1990)). Next, they identify positional orthologs based on the chromosomal positions of homologous gene pairs. When comparing two genomes, G_A and G_B, the MSOAR algorithm (Fu *et al.*, 2007) assigns positional orthologs so as to minimize the number of genome rearrangement

operations required to convert the order of orthologous genes on G_A into that on G_B. The Syntenator algorithm (Rödelsperger and Dieterich, 2008) identifies positional orthologs so as to maximize the size of each conserved gene clusters.

A drawback of those existing algorithms is that they employ a step-wise approach to take into account the conservation of protein sequences and the gene order conservation; they take into account the conservation of protein sequences in the first step, and make use of the gene order conservation in the second step. This step-wise approach requires users to set the threshold of protein sequence conservation score, which would largely affect the results of identification of positional orthologs. Thus, it is desirable to develop a simultaneous approach, which simultaneously takes into account the conservation of protein sequences and the gene order conservation.

In Chapter III, we describe a novel algorithm named OASYS (Orthology Assignment based on SYnteny and Sequence information) (Hachiya and Sakakibara, 2009). OASYS identifies positional orthologs by modeling the probability densities of protein sequence conservation score as well as those of gene order conservation score. Probability densities are estimated for positional orthologs and for other homologous gene pairs, respectively. By making use of these probability densities and applying the decision theory, protein sequence conservation score and gene order conservation score for a homologous gene pair are mapped to the scores in bit-scale. Thus, the two heterogeneous conservation scores can be simply added, or added with weight for each conservation information. Thus, OASYS realizes a simultaneous approach to the identification of positional orthologs based on probabilistic models and the decision theory. As expected, our evaluation tests using prokaryotic genomes demonstrated that OASYS identifies positional orthologs more accurately than existing algorithms, and that OASYS detects conserved gene clusters more sensitively than existing algorithms.

In Chapter III, we also describe a study using OASYS on the relationship between gene order and gene function. We focus on an interesting finding in previous research that the degree of protein sequence conservation of genes in conserved gene clusters is substantially higher than that of the other genes (Dandekar *et al.*, 1998; Lemoine *et al.*, 2007). Although the previous studies do not conduct further analyses for discussing evolutionary forces behind the correlation between protein sequence homology and gene order conservation, we pursue the problem of evolutionary forces by estimating the rate of synonymous substitutions (K_S) and the rate of nonsynonymous substitutions (K_A). The ratio between K_A and K_S (K_A/K_S) can be used to assess how strong evolutionary pressures have enforced conservation of protein sequences because $K_A/K_S = 1$ means neutral mutations, $K_A/K_S < 1$ purifying selections, and $K_A/K_S > 1$ diversifying positive selections (Yang *et al.*, 2000). We can also assess how frequently the coding sequence of a gene has been substituted based on the value of K_S. In this research, we assume that higher degree of protein sequence conservation of genes in conserved gene clusters can be explained by either stronger purifying pressures to maintain protein sequences (lower value of K_A/K_S ratio), lower substitution rate in the coding sequences (lower value of K_S), or both. Based upon this assumption, we aim to unravel which of the three explanations is appropriate for each

taxonomic group. An efficient workflow using OASYS makes it possible to perform a genome-scale systematic comparison of 101 prokaryotic genomes as well as that of 15 fungal genomes. This research gives rise to the following new findings:

- The correlation between protein sequence homology and gene order conservation was observed also in fungal genome comparisons. We firstly reported this finding in our previous paper (Hachiya and Sakakibara, 2009). This finding is important because it suggests that the methodologies to predict protein-protein interactions used in prokaryotic genomes is also useful to predict protein-protein interactions in eukaryotic genomes (or at least fungal genomes).
- Not only stronger purifying pressures to maintain protein sequences but also lower substitution rate of coding sequences induce the correlation between protein sequence homology and gene order conservation. This finding give an impact on the discussion in the previous studies (Dandekar *et al.*, 1998; Huynen *et al.*, 2000; Wolf *et al.*, 2001; Li *et al.*, 2007) because previous studies only proposed stronger purifying pressures.

4 Organization of this dissertation

The remainder of this dissertation is organized as follows:

- In Chapter II, the research on accurate identification of orthologous segments is described.
- In Chapter III, the research on accurate identification of positional orthologs and sensitive detection of conserved gene clusters is described.
- In Chapter IV, summary of this dissertation and further works are described.

Chapter II

Accurate Identification of Orthologous Segments among Multiple Genomes

The accurate detection of orthologous segments (also referred to as syntenic segments) plays a key role in comparative genomics, as it is useful for inferring genome rearrangement scenarios and computing whole genome alignments. Although a number of algorithms for detecting orthologous segments have been proposed, none of them contain a framework for optimizing their parameter values.

In the present study, we propose an algorithm, named OSfinder (Orthologous Segment finder), which uses a novel scoring scheme based on stochastic models (Hachiya *et al.*, 2009). OSfinder takes as input the positions of short homologous regions (also referred to as anchors) and explicitly discriminates orthologous anchors from non-orthologous anchors by using Markov chain models which represent respective geometric distributions of lengths of orthologous and non-orthologous anchors. Such stochastic modeling makes it possible to optimize parameter values by maximizing the likelihood of the input dataset, and to automate the setting of the optimal parameter values.

We validated the accuracies of orthology mapping algorithms on the basis of their consistency with the orthology annotation of genes. Our evaluation tests using mammalian and bacterial genomes demonstrated that OSfinder shows higher accuracy than previous algorithms.

The OSfinder software was implemented as a C++ program. The software is freely available at http://osfinder.dna.bio.keio.ac.jp under the GNU General Public License.

1 Background

The term *orthologous segment* is defined as a set of genomic segments in different organisms descended from a common ancestor without large rearrangements (Dewey *et al.*, 2006). The accurate detection of orthologous segments is essential for the following: inferring rearrangement-based phylogenies (Tesler, 2002; Bourque *et al.*, 2004), reconstructing ancestral genomes (Bourque *et al.*, 2005; Murphy *et al.*, 2005; Ma *et al.*, 2006), computing whole genome alignments (Dewey *et al.*, 2006; Gibbs *et al.*, 2004; Waterston *et al.*, 2002), identifying orthologous genes (Hubbard *et al.*, 2005; Zheng *et al.*, 2005), and detecting non-coding functional elements such as regulatory elements (Frazer *et al.*, 2004). The problem of identifying orthologous segments is referred to as *orthology mapping* (Dewey *et al.*, 2006).

The general strategy of orthology mapping is as follows: (i) Take as input the positions of short homologous regions (also referred to as anchors) detected among the set of genomes under comparison. Homologous genes or bidirectional local sequence

matches are commonly used as anchors. (ii) Detect *collinear* anchors which are distributed in the same order and have the same orientation. (iii) Connect closely located collinear anchors. (iv) Output connected components as orthologous segments.

One difficulty in orthology mapping concerns the fact that a non-negligible fraction of input anchors are non-orthologous rather than orthologous. In the case where the anchors are homologous gene pairs, paralogous gene pairs can be detected as non-orthologous anchors. In the case where the anchors are homologous sequence matches, repeat sequences can be detected as non-orthologous anchors. Conservation scores for anchors and distances between adjacent anchors constitute important features for distinguishing between orthologous genomic regions, in which anchors are distributed densely in off-diagonal positions, and non-orthologous genomic regions, in which anchors are distributed randomly. Existing orthology mapping programs implicitly filter out non-orthologous anchors in the process of identifying orthologous segments. Pevzner and Tesler (2003a) proposed the GRIMM-Synteny algorithm, which chains every pair of anchors if the distance between the two anchors is less than a certain distance threshold, removes chained components if the size of the components is smaller than a certain size threshold, and reports the remaining components as synteny blocks. In order to avoid detecting non-orthologous genomic regions as synteny blocks, it is important to set these two threshold values appropriately. However, GRIMM-Synteny does not provide a framework for determining optimal threshold parameters.

ADHoRe (Vandepoele *et al.*, 2002) and SyMAP (Soderlund *et al.*, 2006) are tools for detecting orthologous segments which are capable of automatically determining the distance threshold value. These tools perform detection by starting with a small value of the distance threshold and increasing it iteratively. This iteration process yields an appropriate distance threshold value which maximizes the length of the orthologous segments while retaining satisfactory quality (Soderlund *et al.*, 2006). Both ADHoRe and SyMAP define the quality of the orthologous segments on the basis of the diagonal properties of the anchor positions. For a series of anchors, the anchor positions are fitted with a linear regression model, and the quality is computed as the coefficient of determination. Although these programs can determine the distance threshold automatically, they require a quality threshold to be set manually.

In addition to the above programs, other orthology mapping algorithms, including DAGChainer (Haas *et al.*, 2004), AXTCHAIN (Kent *et al.*, 2003), DiagHunter (Cannon *et al.*, 2003), FISH (Calabrese *et al.*, 2003), and Cinteny (Sinha and Meller, 2007), also require the manual setting of key threshold parameters. Since these thresholds can affect the accuracy of orthology mapping programs and are difficult to set manually, a more sophisticated approach for determining their parameter values is needed. Furthermore, the vast majority of existing orthology mapping programs are applicable only in pairwise genome comparisons. Thus, the capability to compare multiple genomes is also desired.

In the present study, we propose an orthology mapping algorithm, named OSfinder (Orthologous Segment finder), which uses a novel scoring scheme based on stochastic models. OSfinder explicitly discriminates orthologous anchors from non-orthologous

anchors by using Markov chain models, which represent respective geometric distributions of lengths of orthologous and non-orthologous anchors. Such stochastic modeling makes it possible to optimize parameter values by maximizing the likelihood of the input dataset, and to automate the setting of the optimal parameter values. Moreover, OSfinder can be applied not only in pairwise genome comparisons, but also in multiple genome comparisons. There is no limit to the number of genomes which can be compared with our software.

2 Methods

2.1 Detecting anchors

The term *anchor* generally refers to well-conserved short regions between two or multiple genomes, and is biologically defined as a group of homologous genes or a set of homologous sequence matches. In our experiments, anchors were detected between mammalian genomes and between bacterial genomes. Mammalian genomes included those of human, chimpanzee, macaque, mouse, rat, dog, and opossum, and bacterial genomes included those of *M. tuberculosis* (Mtu), *M. bovis* (Mbo), *M. leprae* (Mle), and *M. avium* (Mpa). Since the method for detecting anchors can affect the accuracy of orthology mapping programs, two methods for detecting anchors were taken into account.

2.1.1 DNA sequence matches

Whole genome sequences of the seven mammals and the four bacteria were taken from the Ensembl genome browser (Hubbard *et al.*, 2007) and the RefSeq database (Pruitt *et al.*, 2007), respectively. When comparing two genomes x and x', the whole genome sequences of x and x' were input into Murasaki (Popendorf *et al.*, 2007) with the repeat mask option. The genomic locations of the anchors were then output by Murasaki. After appropriate format transformation, the Murasaki output was used as input for the orthology mapping programs. Both pairwise and multiple anchors can be computed by this work flow.

2.1.2 Homologous protein sequences

Protein sequences encoded in the seven mammalian genomes and the four bacteria genomes were drawn from the Ensembl genome browser and the RefSeq database, respectively. When comparing two genomes x and x', all protein sequences encoded in genome x were compared with all protein sequences encoded in genome x' by using the BLASTP program (Altschul *et al.*, 1990), and protein pairs whose E-values were less than 10^{-100} were regarded as anchors. Then, pairs of gene IDs were transformed into pairs of genomic locations of the genes. The file containing the genomic positions of the anchors were used as input for the orthology mapping programs. Only pairwise anchors can be computed by this work flow.

The statistics for the anchors detected between mammalian genomes are summa-

Table. II.1 Statistics for anchors in the pairwise comparison of mammalian genomes

Genomes	Number	Length
DNA sequence matches		
human-chimpanzee	3,588,387	386
human-macaque	3,528,953	174
human-mouse	227,258	99
human-rat	205,271	98
human-dog	542,745	107
human-opossum	78,659	108
Homologous gene pairs		
human-chimpanzee	97,865	37,389
human-macaque	93,346	38,288
human-mouse	90,088	39,702
human-rat	64,257	46,096
human-dog	41,020	62,429
human-opossum	104,093	35,024

We show the number and the average length (bp) of anchors detected in the pairwise comparison of mammalian genomes. The lengths shown here were calculated along with the human genome.

rized in Table II.1.

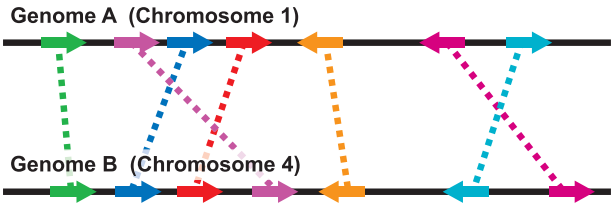
2.2 Mathematical definitions

The OSfinder algorithm is based on the following mathematical definitions. Here, we denote the set of genomes under comparison as G and the set of pre-computed anchors as a.

2.2.1 Anchors

The genomic position of an anchor a_i ($\in a$) can be represented by four properties for each genome x ($\in G$): chromosome ID ($a_i.chrom_x$), start position ($a_i.start_x$), end position ($a_i.end_x$), and strand information ($a_i.sign_x$) (Figs. II.1(a), II.1(b)). We define that the values of $a_i.start_x$ and $a_i.end_x$ are positive integer, and represent coordinates in terms of forward strand positions in the chromosome $a_i.chrom_x$, where the first base in a chromosome is numbered 1. That is, for an anchor a_i on the reverse strand, the start and end positions of a_i are defined as the coordinates of the complementary region of a_i on the forward strand, and therefore these two values satisfy the condition $a_i.start_x < a_i.end_x$ regardless of the value of $a_i.sign_x$. Further, we assume that for the reference genome $\dot{x}$, the value of $a_i.sign_{\dot{x}}$ is "1" $^{\forall}a_i \in a$. The value of $a_i.sign_x$ is "1" if the anchor region from the genome x is not inverted relative to the anchor region from the reference genome $\dot{x}$, and $a_i.sign_x = -1$ if the anchor region from the genome x is inverted relative to the anchor region from the reference

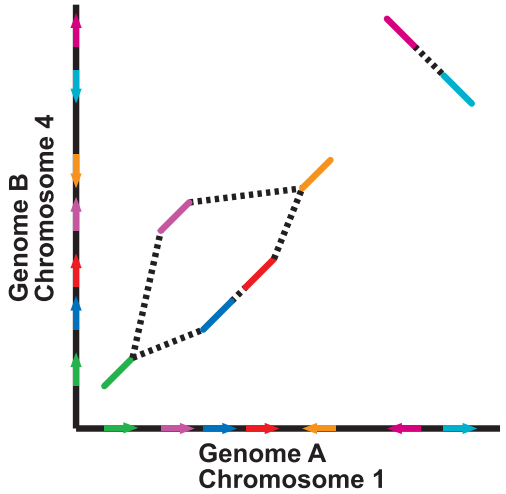
(a) Genomic positions of the pre-computed anchors



(b) Properties of the pre-computed anchors

Genome A				Genome B			
chrom	start	end	sign	chrom	start	end	sign
1	100	200	1	4	150	250	1
1	300	400	1	4	700	800	1
1	450	550	1	4	350	450	1
1	600	700	1	4	500	600	1
1	800	900	1	4	850	900	1
1	1100	1200	1	4	1350	1450	-1
1	1300	1400	1	4	1150	1250	-1

(c) Dot-plot visualization of the anchor graph



(d) Chain extraction

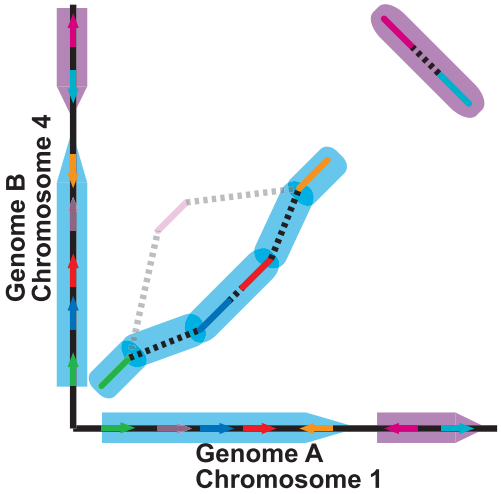


Fig. II.1 These figures show a toy problem of detecting chains in which seven anchors are pre-computed between two genomes A and B. (a) shows the genomic positions of the seven anchors, in which colored arrows represent the anchors (either homologous genes or conserved sequences). In this figure, anchors located on the forward strand are depicted as right arrows, and anchors located on the reverse strand are depicted as left arrows. (b) represents the properties of the seven anchors, in which genome A is used as the reference genome. (c) visualizes the anchor graph by using a dot-plot, in which anchors are depicted by colored solid lines and edges are depicted by black broken lines. (d) illustrates that chains (indicated by colored blocks) correspond to non-intersecting paths in the anchor graph.

genome $\dot{x}$. Note that the choice of the reference genome does not affect the *collinear* relation between the anchors.

2.2.2 Collinearity

In comparative genomics, conservations which are distributed in the same order and have the same orientation are referred to as *collinear* conservations (Bennetzen and Ramakrishna, 2002; Song *et al.*, 2002). Two anchors, a_i and $a_{i'}$, are *collinear* if

the following conditions are satisfied:

$$\begin{aligned} & a_i.chrom_x = a_{i'}.chrom_x \\ & a_i.sign_x = a_{i'}.sign_x \\ & \begin{cases} a_i.end_x < a_{i'}.start_x & \text{when } a_i.sign_x = 1 \\ a_{i'}.end_x < a_i.start_x & \text{when } a_i.sign_x = -1, \end{cases} \end{aligned} \tag{II.1}$$

$^{\forall}x \in G$. Let $a_i \prec a_{i'}$ denote the case where a_i and $a_{i'}$ satisfy the conditions shown in Eq. (II.1).

2.2.3 Anchor graph

The collinear relation defines a partial order between the anchors. Since a partial order induces a directed acyclic graph (DAG), the collinear relations can be represented as a DAG (Fig. II.1(c)). OSfinder constructs a DAG in which a node is an anchor and a directed edge is drawn from a_i to $a_{i'}$ if $a_i \prec a_{i'}$ and there is no anchor $a_{i''}$ satisfying $a_i \prec a_{i''} \prec a_{i'}$. We call this type of DAG *an anchor graph*.

2.2.4 Edges in an anchor graph

The start and end positions of an edge e_j ($\in e$) connecting two anchors (a_i and $a_{i'}$) are defined as follows:

$$\begin{aligned} e_j.start_x &\equiv \min\{a_i.end_x, a_{i'}.end_x\} + 1 \\ e_j.end_x &\equiv \max\{a_i.start_x, a_{i'}.start_x\} - 1. \end{aligned}$$

2.2.5 Length of anchors and edges

The length of an anchor a_i and the length of an edge e_j are defined as follows:

$$\begin{aligned} a_i.length &\equiv \sum_{x \in G} (a_i.end_x - a_i.start_x + 1) \\ e_j.length &\equiv \sum_{x \in G} (e_j.end_x - e_j.start_x + 1). \end{aligned}$$

2.2.6 Chains

Chains are genomic segments in which anchors are distributed densely in off-diagonal positions. Chains correspond exactly to *non-intersecting* suboptimal paths in the observed anchor graph, where two paths are intersecting if their coordinate spans overlap with each other $^{\forall}x \in G$ (Fig. II.1(d)).

2.2.7 Orthologous segments

Orthologous segments are defined as genomic segments descended from a common ancestor without large rearrangements. An orthologous segment corresponds to a sequence of collinear chains, where the collinearity of chains is defined in a similar manner to that of anchors (Fig. II.2).

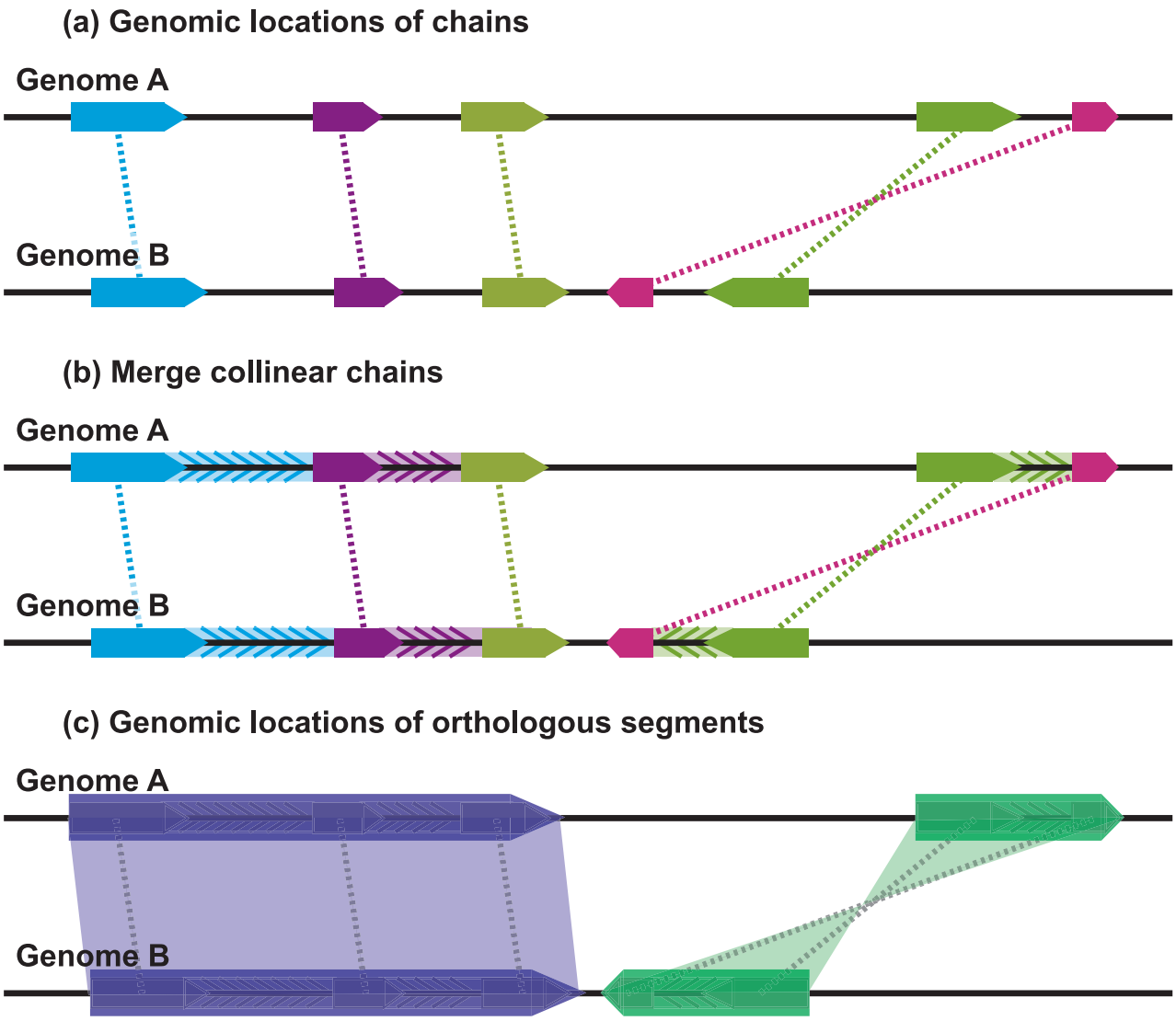


Fig. II.2 These figures show the algorithm which detects orthologous segments. Here, we assume that five chains are detected between two genomes A and B. (a) illustrates the genomic locations of the chains along with the two genomes, where the chains are represented by colored arrows. (b) visualizes the collinear relation between the chains by highlighting the genomic regions between collinear chains. (c) exhibits the genomic locations of the orthologous segments, where orthologous segments are represented by large colored arrows.

2.3 OSfinder algorithm

In order to identify orthologous segments accurately, orthology mapping algorithms should be able to distinguish between orthologous and non-orthologous anchors. For this purpose, OSfinder introduces a set of hidden variables named *labels*. A label is assigned to an anchor or an edge, and its value is either "+" or "−", where "+" represents an orthologous anchor or edge and "−" represents a non-orthologous one.

The likelihood for the observed anchor graph is defined by two sets of variables, namely a set of model parameters M and a set of labels L. By computing the maximum likelihood solution for M and L, the respective length distributions of orthologous and non-orthologous anchors (edges) are fitted to geometric distributions defined by Markov chain models. The optimized model parameters not only determine the optimal length threshold for anchors (edges) which is used to discriminate between orthologous and non-orthologous anchors (edges), but also provide the score for anchors (edges) in the anchor graph. Based on the scores, non-intersecting suboptimal paths are efficiently extracted from the anchor graph by using a dynamic programming technique, and a set of chains is detected. Finally, a sequence of collinear chains is merged into an orthologous segment in order to fill large gap regions between collinear chains.

The overall algorithm of OSfinder is composed of the following steps. (i) Take the genomic positions of the anchors as input. (ii) Construct an anchor graph. The definition of the likelihood for an anchor graph is described in "2.3.1 Likelihood for an anchor graph". (iii) Compute the optimal values for labels and model parameters. Two optimization algorithms are described in "2.3.2 Global maximization algorithm" and "2.3.3 Local maximization algorithm". (iv) Extract non-intersecting suboptimal paths from the observed anchor graph and detect a set of chains. The description of the extraction algorithm can be found in "2.3.4 Chain extraction algorithm". (v) Merge collinear chains and output the merged components as orthologous segments. The merge algorithm is described in "2.3.5 Merge algorithm".

2.3.1 Likelihood for an anchor graph

OSfinder models the respective length distributions of orthologous and non-orthologous anchors (and edges) in the observed anchor graph by using Markov chain models. These models have two states, the extend state (X) and the end state (N), and two state transitions, $X \to X$ and $X \to N$ (Fig. II.3). We denote the transition probability from state X to state X as $P(X \to X|M)$, and the transition probability from state X to state N as $P(X \to N|M)$, where M denotes a model. Note that $P(X \to X|M) + P(X \to N|M) = 1$. An anchor (or an edge) whose length is l indicates the transition sequence $(X \to X)^{l-1} \to N$. Thus, the likelihood for an anchor a_i and the likelihood for an edge e_j are defined by the following geometric distributions:

$$P(a_i|M) = P(X \to X|M)^{(a_i.length-1)} \times P(X \to N|M)$$
$$P(e_j|M) = P(X \to X|M)^{(e_j.length-1)} \times P(X \to N|M).$$

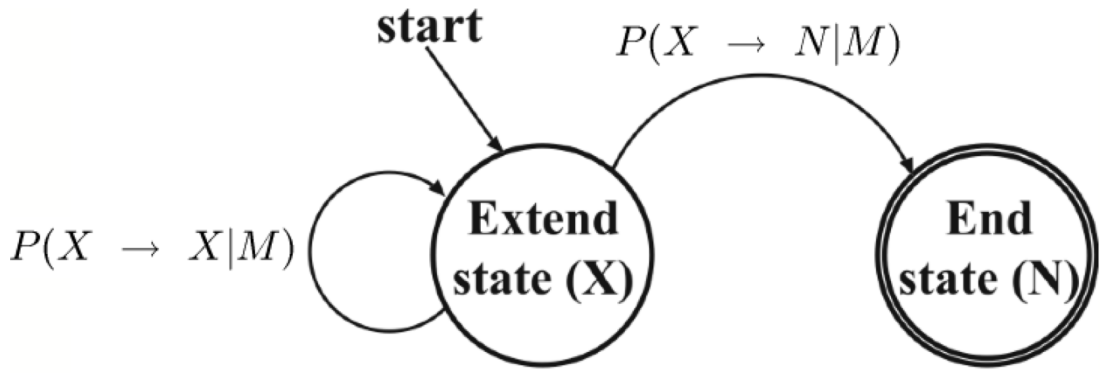


Fig. II.3 Markov chain models in OSfinder.

Next, we set up four Markov chain models, namely a model for representing orthologous anchors (M^{+}_{anchor}), a model for representing non-orthologous anchors (M^{-}_{anchor}), a model for representing orthologous edges (M^{+}_{edge}), and a model for representing non-orthologous edges (M^{-}_{edge}). OSfinder assumes that the average length of orthologous anchors is greater than the average length of non-orthologous anchors. This assumption is implemented in the constraint shown in Eq. (II.2). Similarly, it is assumed that the average length of orthologous edges is shorter than the average length of non-orthologous edges. This assumption is implemented in the constraint shown in Eq. (II.3).

$$P(X \rightarrow N|M^{+}_{\text{anchor}}) < P(X \rightarrow N|M^{-}_{\text{anchor}}) \tag{II.2}$$

$$P(X \rightarrow N|M^{+}_{\text{edge}}) > P(X \rightarrow N|M^{-}_{\text{edge}}). \tag{II.3}$$

Given a set of model parameters M and a set of labels L, OSfinder defines the likelihood for an anchor a_i and the likelihood for an edge e_j as follows:

$$\begin{aligned} P(a_i|M,L) &= \begin{cases} P(a_i|M^{+}_{\text{anchor}}) & \text{if } a_i.label \text{ is ``+''}, \\ P(a_i|M^{-}_{\text{anchor}}) & \text{if } a_i.label \text{ is ``$-$''}. \end{cases} \\ P(e_j|M,L) &= \begin{cases} P(e_j|M^{+}_{\text{edge}}) & \text{if } e_j.label \text{ is ``+''}, \\ P(e_j|M^{-}_{\text{edge}}) & \text{if } e_j.label \text{ is ``$-$''}. \end{cases} \end{aligned} \tag{II.4}$$

Given a set of labels L, let a^{+} be a set of anchors labeled as "+" and a^{-} be a set of anchors labeled as "−" ($a = a^{+} \cup a^{-}$). Similarly, let e^{+} be a set of edges labeled as "+" and e^{-} be a set of edges labeled as "−" ($e = e^{+} \cup e^{-}$). Then, given M and L, the likelihood for an anchor graph $G = (a, e)$ is defined as follows:

$$\begin{aligned} &P(a,e|M,L) \\ &= \Pi_{a_i \in a^{+}} P(a_i|M^{+}_{\text{anchor}}) \times \Pi_{a_{i'} \in a^{-}} P(a_{i'}|M^{-}_{\text{anchor}}) \\ &\quad \times \Pi_{e_j \in e^{+}} P(e_j|M^{+}_{\text{edge}}) \times \Pi_{e_{j'} \in e^{-}} P(e_{j'}|M^{-}_{\text{edge}}). \end{aligned} \tag{II.5}$$

2.3.2 Global maximization algorithm

The parameter values in our Markov chain models are optimized so as to maximize the likelihood for the observed anchor graph. Let $\tilde{M}$ denote a set of *optimal* model parameters and $\tilde{L}$ denote a set of *optimal* labels. $\tilde{M}$ and $\tilde{L}$ are defined by the following equation:

$$(\tilde{M}, \tilde{L}) = \mathrm{argmax}_{(M,L)} P(G|M, L). \tag{II.6}$$

Given a set of labels L, the *conditionally optimal* parameter set $\hat{M}_L$ is defined as follows:

$$\hat{M}_L = \mathrm{argmax}_M P(G|M, L). \tag{II.7}$$

The conditionally optimal model parameters can be calculated from the following equations (see Proof 1 in the section 2.4.1):

$$\begin{aligned}
P(X \to N|M^+_{\mathrm{anchor}}) &= \frac{|a^+|}{\sum_{a_i \in a^+} a_i.length} \\
P(X \to N|M^-_{\mathrm{anchor}}) &= \frac{|a^-|}{\sum_{a_{i'} \in a^-} a_{i'}.length} \\
P(X \to N|M^+_{\mathrm{edge}}) &= \frac{|e^+|}{\sum_{e_j \in e^+} e_j.length} \\
P(X \to N|M^-_{\mathrm{edge}}) &= \frac{|e^-|}{\sum_{e_{j'} \in e^-} e_{j'}.length}.
\end{aligned} \tag{II.8}$$

Note that $P(X \to X|M)$ can be calculated from $P(X \to N|M)$ easily.

From Eq. (II.7), the maximization problem shown in Eq. (II.6) can be restated as follows:

$$\begin{aligned}
\max_{M,L} P(G|M, L) &= \max_L \max_M P(G|M, L) \\
&= \max_L P(G|\hat{M}_L, L).
\end{aligned}$$

Thus, a naive method to maximize the likelihood for an anchor graph is to enumerate all possible label sets and to find the label set which maximizes $P(G|\hat{M}_L, L)$. However, the computation of the naive method is infeasible in terms of computational costs since the number of all possible label sets is $2^{(|a|+|e|)}$. Our global maximization algorithm reduces the number of label sets for enumeration to $(|a| + |e| - 2)$ without losing the ability to find the global optimum. The total computational complexity of the global maximization algorithm is $O(|a|^2 + |e|^2)$.

Let $M_{\mathrm{anchor}} = \{M^+_{\mathrm{anchor}}, M^-_{\mathrm{anchor}}\}$, $M_{\mathrm{edge}} = \{M^+_{\mathrm{edge}}, M^-_{\mathrm{edge}}\}$, L_{anchor} be a set of labels for anchors, and L_{edge} be a set of labels for edges. We define $P(a|M_{\mathrm{anchor}}, L_{\mathrm{anchor}})$

and $P(e|M_{\text{edge}}, L_{\text{edge}})$ as follows:

$$\begin{aligned} &P(a|M_{\text{anchor}}, L_{\text{anchor}}) \\ &\quad = \Pi_{a_i \in a^+} P(a_i|M^+_{\text{anchor}}) \times \Pi_{a_{i'} \in a^-} P(a_{i'}|M^-_{\text{anchor}}) \\ &P(e|M_{\text{edge}}, L_{\text{edge}}) \\ &\quad = \Pi_{e_j \in e^+} P(e_j|M^+_{\text{edge}}) \times \Pi_{e_{j'} \in e^-} P(e_{j'}|M^-_{\text{edge}}). \end{aligned} \tag{II.9}$$

Then, Eq. (II.5) can be restated as Eq. (II.10).

$$P(a, e|M, L) = P(a|M_{\text{anchor}}, L_{\text{anchor}}) \times P(e|M_{\text{edge}}, L_{\text{edge}}). \tag{II.10}$$

From Eq. (II.10), (M_{anchor}, L_{anchor}) and (M_{edge}, L_{edge}) can be optimized separately. Note that $M = M_{\text{anchor}} \cup M_{\text{edge}}$ and $L = L_{\text{anchor}} \cup L_{\text{edge}}$.

Next, we show the algorithm used for the optimization of L_{anchor} and M_{anchor}. We can prove that there exists an *optimal* cutoff value for the anchor length such that the optimal label of an anchor a_i is "+" if the length of a_i is longer than the optimal cutoff length and "−" otherwise (see Proof 2 in the section 2.4.2). Then, the problem of searching the optimal label set for anchors $\tilde{L}_{\text{anchor}}$ is reduced to the problem of searching the optimal cutoff values for the anchor length. Our brute force algorithm enumerates all possible cutoff values and finds the global optimum for L_{anchor} and M_{anchor} as follows:

1. Sort anchors in order of increasing length.
2. For each anchor a_i in the sorted set, $(1 \leq i \leq |a| - 1)$
   (a) Set *cut_length* at $\frac{a_i.length + a_{i+1}.length}{2}$.
   (b) Calculate L_{anchor} on the basis of the current value of *cut_length*.

$$\begin{aligned} a_i.label &= \text{“+”} \quad \text{if } a_i.length > cut\_length \\ a_i.label &= \text{“−”} \quad \text{otherwise.} \end{aligned}$$

   (c) Calculate M_{anchor} by substituting L_{anchor} into Eq. (II.8).
   (d) Calculate $P(a|M_{\text{anchor}}, L_{\text{anchor}})$ by substituting M_{anchor} and L_{anchor} into Eq. (II.9).
3. Report the global optimum solution for M_{anchor} and L_{anchor} which show the highest value of $P(a|M_{\text{anchor}}, L_{\text{anchor}})$. .

The optimization of L_{edge} and M_{edge} can be performed in a manner similar to the optimization of L_{anchor} and M_{anchor}.

2.3.3 Local maximization algorithm

The computation of the global maximization algorithm is also infeasible when the number of anchors or the number of edges is extremely large (Table II.2). Thus, a fast learning algorithm whose computational complexity is $O(|a|+|e|)$ is also implemented in OSfinder.

Table. II.2 Computational time of our local and global maximization algorithms

Genomes	#Anchors[a]	#Edges[b]	Local	Global
DNA sequence matches				
human-chimpanzee	3,588,387	4,987,547	6,550	>10,000
human-macaque	3,528,953	3,963,964	8,787	>10,000
human-mouse	227,258	329,567	17	748
human-rat	205,271	290,624	17	565
human-dog	542,745	766,906	80	4137
human-opossum	78,659	91,006	5	70
Homologous gene pairs				
human-chimpanzee	97,865	2,992,730	221	>10,000
human-macaque	93,346	2,600,844	159	>10,000
human-mouse	90,088	1,711,309	33	>10,000
human-rat	64,257	886,296	17	3,399
human-dog	41,020	206,393	2	200
human-opossum	104,093	2,627,033	74	>10,000

This figure shows the computational time (in min.) of our local and global maximization algorithms. The CPU used in our experiment was 3.0GHz Xeons.

[a]Number of anchors in the anchor graph.

[b]Number of edges in the anchor graph.

Given a set of model parameters M, the *conditionally optimal* label set $\hat{L}_M$ is defined as follows:

$$\hat{L}_M = \mathrm{argmax}_L P(G|M, L). \tag{II.11}$$

The conditionally optimal labels are given by Eq. (II.12).

$$\begin{aligned} a_i.label &= \mathrm{argmax}_{label\in\{+,-\}} P(a_i|M^{label}_{\mathrm{anchor}}) \\ e_j.label &= \mathrm{argmax}_{label\in\{+,-\}} P(e_j|M^{label}_{\mathrm{edge}}) \end{aligned} \tag{II.12}$$

The following algorithm is capable of finding a local optimum.

1. Set the initial model parameters M^0 on the basis of Eq. (II.13)

$$\begin{aligned} P(X \to N|M^{+}_{\mathrm{anchor}}) &= \frac{2}{\mathrm{ave}_{a_i\in a}(a_i.length) + \max_{a_i\in a}(a_i.length)} \\ P(X \to N|M^{-}_{\mathrm{anchor}}) &= \frac{2}{\mathrm{ave}_{a_i\in a}(a_i.length) + \min_{a_i\in a}(a_i.length)} \\ P(X \to N|M^{+}_{\mathrm{edge}}) &= \frac{2}{\mathrm{ave}_{e_j\in e}(e_j.length) + \min_{e_j\in e}(e_j.length)} \\ P(X \to N|M^{-}_{\mathrm{edge}}) &= \frac{2}{\mathrm{ave}_{e_j\in e}(e_j.length) + \max_{e_j\in e}(e_j.length)}. \end{aligned} \tag{II.13}$$

Table. II.3 Accuracy of our local and global maximization algorithms

Genomes	Local			Global		
	Sn	Sp	F	Sn	Sp	F
DNA sequence matches						
human-chimpanzee	98.98	98.38	98.68	-	-	-
human-macaque	97.93	96.86	97.40	-	-	-
human-mouse	93.46	95.18	94.31	93.46	95.18	94.31
human-rat	**89.09**	**90.21**	**89.65**	**89.11**	**90.23**	**89.67**
human-dog	95.70	95.98	95.84	95.70	95.98	95.84
human-opossum	76.43	83.09	79.62	76.43	83.09	79.62
Homologous gene pairs						
human-chimpanzee	98.66	98.41	98.53	-	-	-
human-macaque	90.59	90.79	90.69	-	-	-
human-mouse	86.82	91.43	89.06	-	-	-
human-rat	85.77	87.59	86.67	85.77	87.59	86.67
human-dog	93.20	95.28	94.23	93.20	95.28	94.23
human-opossum	55.92	55.81	55.87	-	-	-

This figure shows the respective accuracies of our local and global maximization algorithms. "–" indicates that OSfinder was unable to calculate orthologous segments within 10,000 minutes. If the respective accuracies of the two methods were different, the values of the accuracies are shown as bold letters.

Parameter values initialized by using Eq. (II.13) clearly satisfy the two conditions shown in Eqs. (II.2) and (II.3).

2. For each step t ($1 \leq t \leq t_{max}$)
   (a) Calculate the conditionally optimal labels by using the parameter values calculated at step $(t-1)$. In other words, $L^t = \hat{L}_{M^{(t-1)}}$.
   (b) Calculate the conditionally optimal model parameters by using the labels calculated at step t. In other words, $M^t = \hat{M}_{L^t}$
   (c) Stop the iteration if $M^{(t-1)} = M^t$.
3. Report the parameter values obtained at the end of the above iteration.

It can be proven that the parameters identified by the algorithm locally maximize the likelihood for the observed anchor graph (see Proof 3 in the section 2.4.3). The default value for $t_{\max}$ is set at 100 in the current version of OSfinder. Although the fast algorithm calculates a local optimum rather than the global optimum, our computational experiments using mammalian genomes show that the accuracy of the local maximization algorithm is almost the same as that of the global maximization algorithm (Table II.3).

2.3.4 Chain extraction algorithm

Given a set of optimal model parameters $\tilde{M}$, the score for an anchor a_i and the score for an edge e_j are defined as log-odds of two likelihoods as follows:

$$a_i.score = \log \frac{P(a_i|M^+_{\text{anchor}})}{P(a_i|M^-_{\text{anchor}})} \tag{II.14}$$

$$e_j.score = \log \frac{P(e_j|M^+_{\text{edge}})}{P(e_j|M^-_{\text{edge}})}. \tag{II.15}$$

A path in an anchor graph corresponds exactly to a sequence of collinear anchors. Let a^{p_k} denote a set of anchors included in a path p_k, and e^{p_k} denote a set of edges included in p_k. In this case, the score for path p_k is defined by the following equation:

$$p_k.score = \sum_{a_i \in a^{p_k}} a_i.score + \sum_{e_j \in e^{p_k}} e_j.score.$$

The path with the highest score can be efficiently found by using a dynamic programming technique with the following recursive formula for the best path ending at anchor a_i:

$$\begin{aligned} &path\_score(a_i) \\ &= a_i.score + \max \begin{cases} \max_{a_{i'} \prec a_i} \{path\_score(a_{i'}) + e_{a_{i'} \to a_i}.score\} \\ 0 \end{cases}, \end{aligned} \tag{II.16}$$

where $e_{a_{i'} \to a_i}$ represents the edge drawn from anchor $a_{i'}$ to anchor a_i. After the optimization of the model parameters, the chain extraction algorithm in OSfinder detects non-intersecting suboptimal paths from the observed anchor graph. It recursively executes the following operations: (i) calculation of the list whose i-th element deposits the value of $path\_score(a_i)$ defined by Eq. (II.16), (ii) detection of the highest-scoring path by tracing back from the element which has the highest value of $path\_score(a_i)$, and (iii) removal of the anchors and edges which are intersecting with the extracted path, until there are no paths scoring higher than zero (Fig. II.4). The re-calculation of the list is essential for the detection of the next highest-scoring path because the removal of the anchors and edges changes the scores in the list. We call the suboptimal paths *chains*.

2.3.5 Merge algorithm

Given a set of chains, the merge algorithm in OSfinder performs the following operation on the basis of a user-defined parameter named *minimum segment length*:

1. Construct a DAG, where a node is a chain and a directed edge is drawn from a chain c_l to a chain $c_{l'}$ if $c_l \prec c_{l'}$ and there is no chain $c_{l''}$ satisfying $c_l \prec c_{l''} \prec c_{l'}$. We call the DAG *a chain graph*.

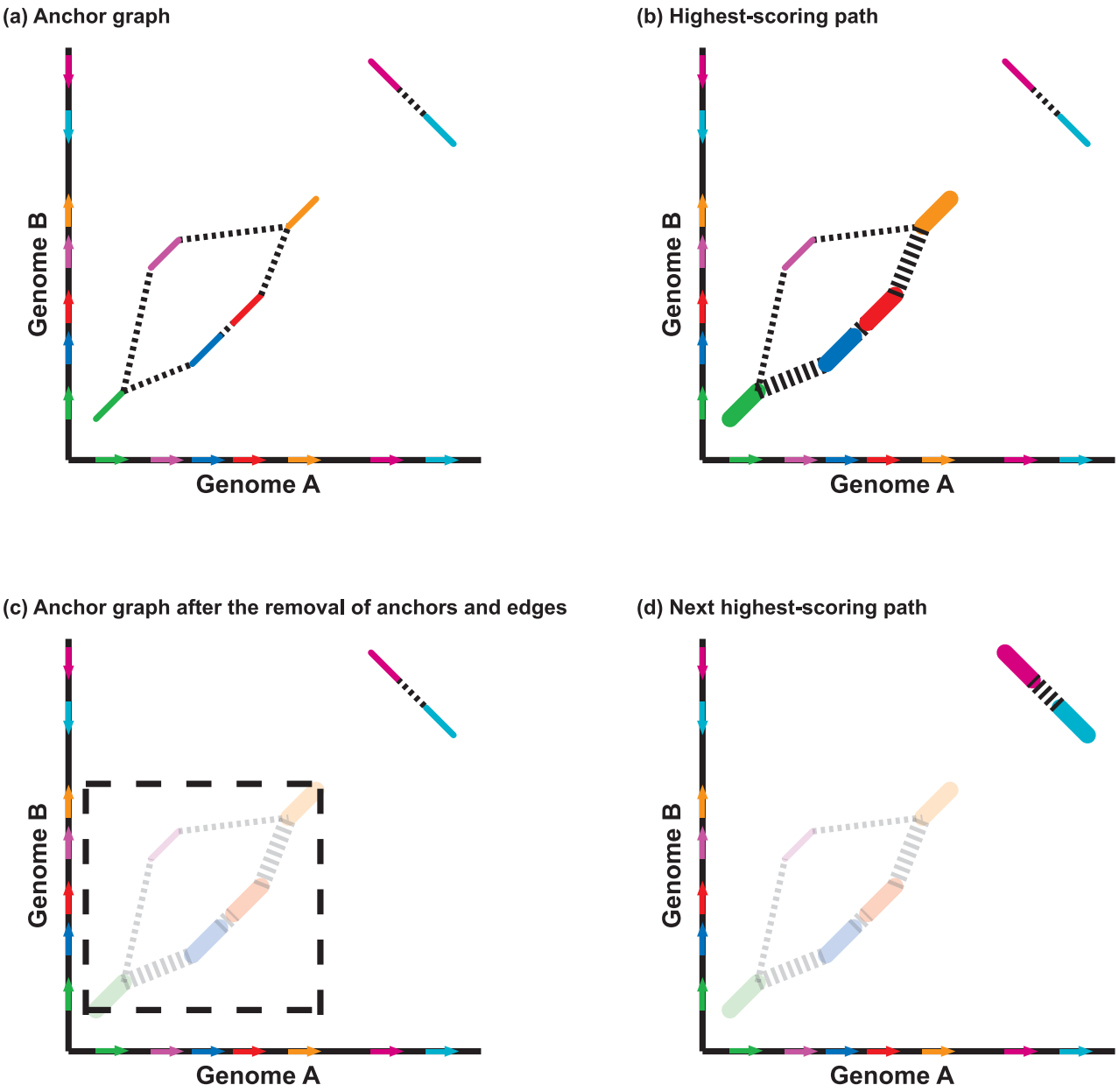


Fig. II.4 These figures show the algorithm which extracts chains. (a) shows the observed anchor graph. Given the anchor graph, (b) illustrates the highest-scoring path, where the anchors and edges included in the highest-scoring path are represented by bold lines. Furthermore, (c) shows the anchor graph after the removal of the anchors and edges which are intersecting with the extracted path. Also, (d) exhibits the next highest-scoring path, where the anchors and edges included in the next highest-scoring path are represented by bold lines.

2. Sort the edges in the chain graph in order of increasing edge length.
3. For each edge e_j in the sorted order $(1 \leq j \leq |e|)$
   (a) Check whether there exists any merged component such that its coordinate

span overlaps with the coordinate span of the edge e_j for at least one genome x ($\in G$), and its length is greater than the *minimum segment length*.

(b) If no such case exists, merge chains connected through the j-th edge.

4. Report merged components whose length is greater than the *minimum segment length* as orthologous segments.

The *minimum segment length* controls the resolution of the orthology mapping. A large value for this parameter is appropriate for analyzing macrorearrangements, and a small value for the parameter is appropriate for drawing detailed dot-plots.

2.4 Proofs

In this section, we use the following representations.

G	: an anchor graph.
a^+	: a set of anchors labeled as "+".
a^-	: a set of anchors labeled as "−".
e^+	: a set of edges labeled as "+".
e^-	: a set of edges labeled as "−".
M^+_{anchor}	: Markov chain model for orthologous anchors.
M^-_{anchor}	: Markov chain model for non-orthologous anchors.
M^+_{edge}	: Markov chain model for orthologous edges.
M^-_{edge}	: Markov chain model for non-orthologous edges.
M	$= \{ M^+_{\text{anchor}}, M^-_{\text{anchor}}, M^+_{\text{edge}}, M^-_{\text{edge}} \}$
L	: a set of labels for anchors and edges.

2.4.1 Proof 1

Here, we demonstrate that the parameter values calculated by using Eq. (II.8) are conditionally optimal.

Note that an anchor (or an edge) whose length is l implies the transition sequence $(X \rightarrow X)^{l-1} \rightarrow N$. Let $N_{a_i}(X \rightarrow X)$ be the number of transitions from state X to state X, which are observed in the transition sequence associated with an anchor a_i. We define $N_{a_i}(X \rightarrow N)$, $N_{e_j}(X \rightarrow X)$ and $N_{e_j}(X \rightarrow N)$ similarly, where e_j represents an edge. The numbers of transitions can be easily calculated by using Eq. (II.17).

$$\begin{aligned} N_{a_i}(X \rightarrow X) &= a_i.length - 1 \\ N_{a_i}(X \rightarrow N) &= 1 \\ N_{e_j}(X \rightarrow X) &= e_j.length - 1 \\ N_{e_j}(X \rightarrow N) &= 1. \end{aligned} \tag{II.17}$$

Given a set of labels for anchors and edges, the maximum likelihood solution for

model parameters is given by Eq. (II.18).

$$
\begin{aligned}
&P(X \to N|M^{+}_{\text{anchor}}) \\
&= \frac{\sum_{a_i \in a^+} N_{a_i}(X \to N)}{\sum_{a_i \in a^+} N_{a_i}(X \to X) + \sum_{a_i \in a^+} N_{a_i}(X \to N)} \\
&P(X \to N|M^{-}_{\text{anchor}}) \\
&= \frac{\sum_{a_{i'} \in a^-} N_{a_{i'}}(X \to N)}{\sum_{a_{i'} \in a^-} N_{a_{i'}}(X \to X) + \sum_{a_{i'} \in a^-} N_{a_{i'}}(X \to N)} \\
&P(X \to N|M^{+}_{\text{edge}}) \\
&= \frac{\sum_{e_j \in e^+} N_{e_j}(X \to N)}{\sum_{e_j \in e^+} N_{e_j}(X \to X) + \sum_{e_j \in e^+} N_{e_j}(X \to N)} \\
&P(X \to N|M^{-}_{\text{edge}}) \\
&= \frac{\sum_{e_{j'} \in e^-} N_{e_{j'}}(X \to N)}{\sum_{e_{j'} \in e^-} N_{e_{j'}}(X \to X) + \sum_{e_{j'} \in e^-} N_{e_{j'}}(X \to N)}.
\end{aligned}
\tag{II.18}
$$

By substituting Eq. (II.17) into Eq. (II.18), Eq. (II.8) is obtained. Thus, the parameter values calculated by using Eq. (II.8) are conditionally optimal.

□

2.4.2 Proof 2

The two conditions shown in Eqs. (II.2) and (II.3) are always satisfied throughout all procedures in OSfinder. Here, we demonstrate that under such conditions, there exists an optimal cutoff value for the anchor length such that the optimal label of an anchor a_i is "+" if the length of a_i is longer than the optimal cutoff length and "−" otherwise.

We assume that the label of an anchor a_i is "+" and the label of an anchor $a_{i'}$ is "−". Then, from Eq. (II.12), Eqs. (II.19) and (II.20) must be satisfied.

$$P(a_i|M^{+}_{\text{anchor}}) > P(a_i|M^{-}_{\text{anchor}}) \tag{II.19}$$

$$P(a_{i'}|M^{+}_{\text{anchor}}) < P(a_{i'}|M^{-}_{\text{anchor}}). \tag{II.20}$$

Let $\Delta\, l = a_{i'}.length - a_i.length$. Assuming $\Delta\, l \geq 0$, the following inequality is

obtained.

$$
\begin{aligned}
&P(a_{i'}|M^{+}_{\text{anchor}}) \\
&= P(a_i|M^{+}_{\text{anchor}}) \times P(X \to X|M^{+}_{\text{anchor}})^{\Delta l} \\
&> P(a_i|M^{-}_{\text{anchor}}) \times P(X \to X|M^{-}_{\text{anchor}})^{\Delta l} \\
&\quad (\because \text{Eqs. (II.19) and (II.2)}) \\
&= P(a_{i'}|M^{-}_{\text{anchor}}).
\end{aligned}
$$

The above inequality conflicts with Eq. (II.20). Thus, $\Delta\, l$ must be negative ($\Delta\, l < 0$). In other words, the length of a_i must be greater than the length of $a_{i'}$. Therefore, it is demonstrated that any anchor labeled as "+" is longer than any anchor labeled as "−". Thus, there exists an optimal cutoff value for the anchor length such that the optimal label of an anchor a_i is "+" if the length of a_i is longer than the optimal cut-off length and"−" otherwise.

□

2.4.3 Proof 3

Here, we demonstrate that parameter values determined by applying our local maximization algorithm are a set of local optimum solution.

Let M^t be model parameters at step t and L^t be labels at step t. Given $M^{(t-1)}$, the labels at step t are calculated from $L^t = \hat{L}_{M^{(t-1)}}$, where $\hat{L}_{M^{(t-1)}}$ is a set of conditionally optimal labels. From Eq. (II.11), the following inequality is obtained:

$$P(G|M^{(t-1)}, L^t) \geq P(G|M^{(t-1)}, L^{(t-1)}), \tag{II.21}$$

where the equality is satisfied if $L^{(t-1)} = L^t$.

Given L^t, the model parameters at step t are calculated from $M^t = \hat{M}_{L^t}$, where $\hat{M}_{L^t}$ is a set of conditionally optimal model parameters. From Eq. (II.7), the following inequality is satisfied:

$$P(G|M^t, L^t) \geq P(G|M^{(t-1)}, L^t), \tag{II.22}$$

where the equality is satisfied iff $M^{(t-1)} = M^t$.

By combining Eq. (II.21) and Eq. (II.22), Eq. (II.23) is obtained.

$$P(G|M^t, L^t) \geq P(G|M^{(t-1)}, L^{(t-1)}), \tag{II.23}$$

where the equality is satisfied if $M^{(t-1)} = M^t$ and $L^{(t-1)} = L^t$. Eq. (II.23) ensures that the likelihood for an anchor graph increases together with the iteration steps until M and L converge. Thus, parameter values determined by applying our local maximization algorithm are a set of local optimum solution.

□

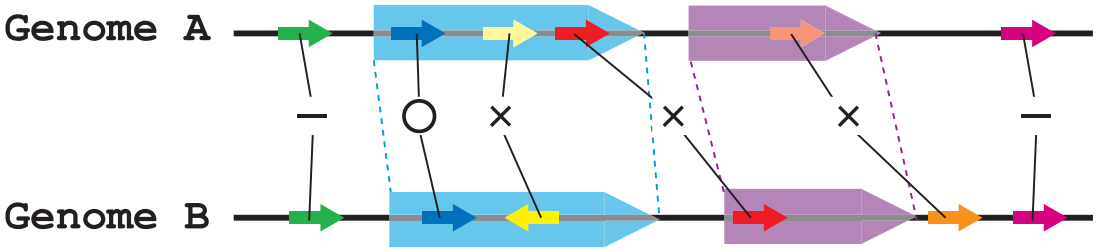


Fig. II.5 Consistency and inconsistency between orthologous gene groups and orthologous segments. This figure shows the respective genomic locations of six pairs of orthologous genes (indicated by small arrows) and two pairs of orthologous segments (indicated by large arrows). The color of the arrows represents the orthologous relationship where pairs of orthologous genes or segments have the same color. "○" ("×") indicates that the gene pair is consistent (inconsistent) with a certain orthologous segment, and "–" indicates that the gene pair is neither consistent nor inconsistent with any orthologous segment.

2.5 Evaluation criteria

Since it is impossible to observe the course of evolutionary history, the evaluation of orthology mapping programs is restricted to simulation experiments (Calabrese *et al.*, 2003; Cannon *et al.*, 2003; Hampson *et al.*, 2003) or assessment on the basis of the consistency of the target program with other orthology mapping programs (Cannon *et al.*, 2003). However, simulation experiments require the mutation models to generate virtual evolutionary histories, and therefore the evaluation results are inevitably biased with respect to the mutation models used in the experiment. In addition, examining the consistency with other programs is not an efficacious methodology for estimating the accuracy if the compared programs are based on similar approaches.

In this chapter, we estimate the accuracy of orthology mapping programs on the basis of their consistency with the orthology annotation of genes (Fig. II.5). Let G be the set of genomes under comparison, $s = \{s_1, s_2, \ldots, s_{|G|}\}$ be an orthologous segment (a set of segments from different genomes), and $g = \{g_1, g_2, \ldots, g_{|G|}\}$ be an orthologous gene group (a set of genes from different genomes), where s_x (g_x) is a segment (gene) from a genome x. Here, we assume that orthologous gene groups do not contain in-paralogs (Remm *et al.*, 2001; Koonin, 2005) and that orthologous relationships are necessarily one-to-one. Then, we define that g is *consistent* with s if, for all $x \in G$, the coding region of g_x overlaps with the coordinate span of s_x and the orientation of g_x is the same as that of s_x. We also define that g is *inconsistent* with s if g is not consistent with s and there exists a genome x in which the coding region of g_x overlaps with the coordinate span of s_x.

Given a set of orthologous gene groups and a set of orthologous segments, let N be the number of orthologous gene groups and N_C be the number of orthologous gene groups with which at least one orthologous segment is consistent. Let $C(s^i)$ denote the number of orthologous gene groups which are consistent with an orthologous segment

s^i and $I(s^i)$ denote the number of orthologous gene groups which are inconsistent with an orthologous segment s^i. Then, the *sensitivity* and the *specificity* are defined as follows:

$$sensitivity = \frac{N_C}{N}$$
$$specificity = \frac{\sum_i C(s^i)}{\sum_i C(s^i) + \sum_i I(s^i)}$$

The F-score is defined as $\frac{2pr}{p+r}$, where p represents the specificity and r represents the sensitivity. We use the F-score as an indicator of accuracy.

In this chapter, the mammalian orthologous gene database SPEED (Vallender *et al.*, 2006) was employed as a reliable source of orthology annotation data for mammalian genes. Regarding the evaluation of the results of bacterial genome comparisons, orthologous gene pairs were identified on the basis of BLAST reciprocal best hits (Tatusov *et al.*, 1997, 2003). Orthologous groups of bacterial genes were calculated by using the method described in (Vallender *et al.*, 2006).

3 Results

3.1 Accuracy in pairwise genome comparisons

3.1.1 Comparison with DAGChainer and ADHoRe

Anchors detected between pairwise genomes were input into DAGChainer (Haas *et al.*, 2004), ADHoRe (Vandepoele *et al.*, 2002), and OSfinder, where DAGChainer and ADHoRe were executed not only with the default parameter values, but also with optimized parameter values. The optimization for the DAGChainer and ADHoRe parameters was performed on the basis of a grid search in order to maximize the F-score. Two parameters in DAGChainer, the average expected distance between two orthologous anchors (-g option) and the maximum allowed distance between two anchors (-D option), were optimized. Two parameters in ADHoRe, the minimum r^2 value (r2_cutoff option) and the maximum distance between the anchors (max_dist option) were optimized. Since OSfinder automatically optimizes its parameter values, there was no need to perform grid searches.

Table II.4 shows the accuracy of the three programs in the pairwise comparison of mammalian genomes. It is worth noting that OSfinder consistently achieved high F-scores (>85% on average), regardless of the anchor type. DAGChainer exhibited low F-scores when the anchors were homologous gene pairs (62.2% with a grid search), while ADHoRe exhibited extremely low F-scores both when the anchors were homologous sequences (47.1% with the default parameters) and when they were homologous gene pairs (39.3% with a grid search). We discuss the reason for the low F-scores achieved by ADHoRe when comparing mammalian genomes in the section "Discussion and Conclusion". These results demonstrate that OSfinder has greater accuracy than the other two programs. The average F-scores of OSfinder were notably higher

than those of DAGChainer and ADHoRe, even though the latter two were executed with optimized parameter values.

The high accuracy of OSfinder is supported further by the results in the pairwise comparison of bacterial genomes (Table II.5). When the anchors were homologous sequences, the average F-score of OSfinder was 85.3%, which was 14.2% higher than that of DAGChainer with a grid search optimization and 20.8% higher than that of ADHoRe with a grid search optimization. When the anchors were homologous gene pairs, the average F-score of OSfinder was 92.2%, which was 7.5% higher than that of DAGChainer with a grid search optimization and 56.2% higher than that of ADHoRe with a grid search optimization.

Table. II.4 Accuracy in the pairwise comparison of mammalian genomes

Genomes	DAGChainer						ADHoRe						OSfinder		
	default			grid search			default			grid Search			default		
	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F
DNA sequence matches															
human-chimpanzee	99.5	73.1	84.3	-	-	-	-	-	-	-	-	-	99.0	98.4	98.7
human-macaque	99.1	77.1	86.7	-	-	-	-	-	-	-	-	-	97.9	96.9	97.4
human-mouse	91.0	88.7	89.8	-	-	-	47.7	46.5	47.1	-	-	-	93.5	95.2	94.3
human-rat	89.2	89.4	89.3	-	-	-	44.8	44.0	44.4	-	-	-	89.1	90.2	89.6
human-dog	97.3	81.0	88.4	-	-	-	49.9	42.3	45.8	-	-	-	95.7	96.0	95.8
human-opossum	50.8	93.8	66.0	-	-	-	48.4	53.8	50.9	-	-	-	76.4	83.1	79.6
Average[a]	82.1	88.2	83.4	-	-	-	47.7	46.7	47.1	-	-	-	88.7	91.1	89.8
Homologous gene pairs															
human-chimpanzee	59.3	26.0	36.2	89.8	28.6	43.4	49.0	47.2	48.1	58.0	54.7	56.3	98.7	98.4	98.5
human-macaque	54.9	30.0	38.8	87.2	31.2	46.0	37.9	35.9	36.8	46.4	42.2	44.2	90.6	90.8	90.7
human-mouse	51.9	47.0	49.3	82.5	49.9	62.2	30.6	27.9	29.2	36.8	34.3	35.5	86.8	91.4	89.1
human-rat	47.2	55.5	51.0	77.4	61.0	68.2	30.6	26.6	28.4	36.4	31.2	33.6	85.8	87.6	86.7
human-dog	52.1	84.8	64.5	88.7	82.1	85.3	38.1	36.6	37.3	43.4	42.4	42.9	93.2	95.3	94.2
human-opossum	31.2	78.3	44.6	70.3	65.6	67.9	24.2	19.1	21.2	27.3	20.4	23.3	55.9	55.8	55.9
Average	49.4	53.6	47.4	82.7	53.1	62.2	35.1	32.2	33.5	41.4	37.5	39.3	85.2	86.6	85.8

We show the respective accuracies of DAGChainer, ADHoRe, and OSfinder in the pairwise comparisons of mammalian genomes. Here, "–" indicates that the calculation of orthologous segments was impossible in our environment due to either time or space limitations.

[a]For the purpose of a fair comparison, the respective accuracies in the human-chimpanzee and human-macaque comparisons were not used in calculating the average values.

Table. II.5 Accuracy results in the pairwise comparison of bacterial genomes

Genomes	DAGChainer						ADHoRe						OSfinder		
	default			grid search			default			grid Search			$S = 10,000^a$		
	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F
DNA sequence matches															
Mtu-Mbo	99.3	16.1	27.8	95.8	35.1	51.4	59.7	88.0	71.1	87.1	83.4	85.2	90.8	99.8	95.1
Mtu-Mle	91.9	39.6	55.3	90.6	68.7	78.1	38.7	41.3	39.9	38.1	42.5	40.2	68.2	76.6	72.1
Mtu-Mpa	92.1	44.6	60.1	91.4	77.1	83.7	69.3	66.6	67.9	70.4	65.9	68.1	91.0	86.3	88.6
Average	94.4	33.4	47.7	92.6	60.3	71.1	55.9	65.3	59.6	65.2	63.9	64.5	83.3	87.6	85.3
Homologous gene pairs															
Mtu-Mbo	99.7	53.4	69.5	99.7	91.4	95.4	39.0	38.5	38.8	98.0	97.9	98.0	99.8	99.8	99.8
Mtu-Mle	89.8	34.0	49.4	86.4	68.4	76.4	16.8	1.5	2.7	28.2	2.8	5.1	84.8	92.6	88.5
Mtu-Mpa	92.4	44.1	59.7	90.8	75.1	82.2	30.7	1.1	2.3	50.5	2.6	4.9	90.5	86.0	88.2
Average	94.0	43.8	59.5	92.3	83.5	84.7	28.8	13.7	14.6	58.9	34.4	36.0	91.7	92.8	92.2

We show the respective accuracies of DAGChainer, ADHoRe, and OSfinder in the pairwise comparison of bacterial genomes.

[a]The value of the *minimum segment length* parameter (S) was set at 10,000bp.

3.1.2 Comparison with syntenic nets

The UCSC genome browser provides a common repository for genomic annotation data (Kuhn *et al.*, 2007; Karolchik *et al.*, 2008). Syntenic nets, which are unique annotations in the UCSC genome browser, are genomic regions descended from a single genomic segment in a common ancestor without macrorearrangements. In Table II.6, we present the accuracy of syntenic nets together with the accuracy of OSfinder. For the purpose of a fair comparison, Table II.6 displays the accuracy of OSfinder when homologous sequences were used as anchors.

We can see in Table II.6 the trade-off between the sensitivity and the specificity. OSfinder achieved a 25.6% higher average specificity than syntenic nets, while the average sensitivity of syntenic nets was 3.0% higher than that of OSfinder. Regarding the average F-score, OSfinder achieved a 14.0% higher value than syntenic nets.

Table. II.6 Accuracy of syntenic nets and OSfinder in the pairwise comparison of mammalian genomes

	Syntenic nets			OSfinder		
Genomes	Sn	Sp	F	Sn	Sp	F
human-chimpanzee	98.9	85.6	91.8	99.0	98.4	98.7
human-macaque	98.5	66.5	79.4	97.9	96.9	97.4
human-mouse	97.2	69.5	81.0	93.5	95.1	94.3
human-rat	97.1	66.3	78.8	89.1	90.2	89.6
human-dog	98.2	60.7	75.0	95.7	96.0	95.8
Average	98.0	69.7	81.2	95.0	95.3	95.2

3.2 Accuracy in multiple genome comparisons

The accuracy of OSfinder in multiple genome comparisons was compared with that of the TBA program (Blanchette *et al.*, 2004) and Mercator (Dewey *et al.*, 2006; Dewey, 2007). The respective accuracies of these programs were evaluated in comparisons of mammalian X chromosomes. For the generation of input for OSfinder, anchors among multiple genomes were detected by using Murasaki (Popendorf *et al.*, 2007). For the generation of input for the TBA program, the BLASTZ alignments (Altschul *et al.*, 1997; Schwartz *et al.*, 2003) between all pairs of genomes under comparison were calculated. For the generation of input for Mercator, homologous gene pairs were detected between all pairs of genomes under comparison by using BLASTP program. Note that TBA and Mercator take as input sets of anchors detected between all pairs of genomes under comparison, whereas OSfinder takes as input a set of anchors detected among multiple genomes. Thus, TBA and Mercator require the $O(N^2)$ iterations of the calculation of anchors, where N represents the number of genomes under comparison, whereas it is sufficient to perform the calculation of anchors among multiple genomes only once for the input of OSfinder.

3.2.1 Procedures for the execution of the TBA program

BLASTZ alignments (Altschul *et al.*, 1997; Schwartz *et al.*, 2003) were calculated between all pairs of genomes under comparison. The parameter values were set at "H=2000 Y=3400 B=2 C=0" (Margulies *et al.*, 2007). After performing the appropriate post-processing, we executed the TBA program (Blanchette *et al.*, 2004) with the phylogenetic tree "((((human chimpanzee)macaque)mouse)dog)" (Nikolaev *et al.*, 2007; Lunter, 2007a).

3.2.2 Procedures for the execution of Mercator

Protein sequences of mammalian genes were drawn from the Ensembl genome browser. For each genome under comparison, the genomic locations of protein-coding genes were included in the *anchors* file. For each pair of genomes under comparison, all pairs of homologous gene IDs detected by the BLASTP program were included in the *hit* file together with bit scores and E-values. Then, Mercator (Dewey *et al.*, 2006; Dewey, 2007) was executed with the default parameters.

3.2.3 Comparison with TBA and Mercator

Table II.7 shows the accuracy in the comparison of multiple mammalian genomes. The results contain two important points. The first is that the accuracy of OSfinder was extremely high, with F-scores of over 95%. The average F-score of OSfinder was 96.9%, which was notably higher than that of TBA (64.2%) and Mercator (91.2%). The second point is that the F-scores of OSfinder in multiple genome comparisons were slightly higher than that in pairwise genome comparisons (Table II.6). For example, the F-score in the human-chimpanzee-macaque-mouse comparison (97.5%) was higher than that in the human-mouse comparison (95.8%). This tendency is also visible in the results for the human-chimpanzee-macaque-dog comparison. These results imply that the accuracy of OSfinder in pairwise comparisons can be improved by adding closely related genome(s) and by comparing multiple genomes.

In Table II.8, we present the accuracy in the comparison of multiple bacterial genomes. These results also demonstrate the high accuracy of OSfinder with F-scores over 90%. Table II.5 and Table II.8 show that the average F-score in the Mtu-Mbo-Mle (Mtu-Mbo-Mpa) comparison was higher than that in the Mtu-Mle (Mtu-Mbo) comparison. The results were the same as the results obtained from the comparison of multiple mammalian genomes.

4 Discussion and Conclusion

The results in this chapter have demonstrated the potential of stochastic models and learning algorithms in OSfinder to improve the accuracy of orthology mapping in both pairwise and multiple genome comparisons. We have shown that our novel algorithm makes it possible to identify orthologous segments with accuracy which is consistently higher than that of other algorithms, without any manual effort to

Table. II.7 Accuracy in the comparison of multiple mammalian genomes

Genomes	TBA			Mercator			OSfinder		
	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F
Hsa-Ptr-Mul[a]	96.7	82.7	89.1	97.6	97.6	97.6	97.6	97.6	97.6
Hsa-Ptr-Mul-Mmu[b]	68.1	43.5	53.1	90.2	85.9	88.0	97.1	97.8	97.5
Hsa-Ptr-Mul-Cfa[c]	96.5	55.0	70.1	95.8	96.9	96.3	96.9	96.9	96.9
Hsa-Ptr-Mul-Mmu-Cfa[d]	66.8	33.1	44.3	85.0	81.2	83.0	94.0	97.5	95.7
Average	82.0	53.6	64.2	92.2	90.4	91.2	96.4	97.5	96.9

[a]Comparison of human, chimpanzee, and macaque genomes.

[b]Comparison of human, chimpanzee, macaque, and mouse genomes.

[c]Comparison of human, chimpanzee, macaque, and dog genomes.

[d]Comparison of human, chimpanzee, macaque, mouse, and dog genomes.

Table. II.8 Accuracy results in the comparison of multiple bacterial genomes

Genomes	TBA			OSfinder		
	Sn	Sp	F	Sn	Sp	F
Mtu-Mbo-Mle	42.8	25.8	32.2	86.6	97.6	91.8
Mtu-Mbo-Mpa	71.9	54.3	61.9	90.6	89.5	90.1
Average	57.4	40.1	47.1	88.6	93.6	91.0

determine the parameter values.

Quality-based methods, such as ADHoRe (Vandepoele *et al.*, 2002) and SyMAP (Soderlund *et al.*, 2006), estimate the quality by computing the coefficient of determination. ADHoRe with a grid search optimization showed an extremely high accuracy in the Mtu-Mbo comparison (98.0% in F-score when the anchors were homologous gene pairs), although its accuracy in the Mtu-Mle and Mtu-Mpa comparisons was profoundly low (5.1% and 4.9% in F-score, respectively). Moreover, ADHoRe also showed low F-scores in mammalian genome comparisons (47.1% in average F-score when the anchors were homologous sequences). These results imply that although quality-based methods are excellent approaches when comparing very closely related genomes, these methods are not adequate when comparing distantly related genomes where the positions of the orthologous anchors can not be fitted with linear regression models.

DAGChainer (Haas *et al.*, 2004) measures the diagonal properties of the input anchors by utilizing a scoring scheme which is more relaxed than linear regression models. The scoring scheme makes it possible to compare distantly related genomes while maintaining a relatively high F-score (about 80% when the anchors are homologous gene pairs in the pairwise comparison of bacterial genomes). The scoring scheme, however, suffers from low specificity when a large fraction of the input anchors are

non-orthologous (e.g., when the anchors are homologous sequences in the pairwise comparison of bacterial genomes). Therefore, the accuracy of distinguishing between orthologous anchors and non-orthologous anchors is the key to performing orthology mapping with consistently high accuracy.

The scoring scheme of OSfinder takes into account the distance between collinear anchors instead of the coefficient of determination of the linear regression. Furthermore, stochastic models are employed in OSfinder for accurately distinguishing between orthologous anchors and non-orthologous anchors. Thus, the OSfinder algorithm consistently achieves high accuracy even when distantly related genomes are compared and when a large fraction of the anchors are not orthologous.

With the rapidly increasing amount of sequence data, the automation of orthology mapping and the ability to compare multiple genomes will continue to become ever more important for high-throughput genome analysis. In addition to the seven mammalian genomic sequences used in our analysis, draft sequences for orangutan (*Pongo pygmaeus abelii*), cow (*Bos taurus*), and horse (*Equus caballus*) have already been made available in the Ensembl genome browser. Furthermore, it is expected that over 20 mammalian genome sequences will become available in the near future. It is expected that the calculation results of OSfinder can be further improved by using the increasing number of closely related genome sequences.

Chapter III

Correlation between Protein Sequence Homology and Gene Order Conservation

A conserved gene cluster (also referred to as a conserved gene order) is defined as a cluster of neighboring genes whose gene order is conserved across several species. In the present study, we propose a novel workflow which enables sensitive detection of conserved gene clusters by taking into account the information of gene order conservation in the step to identify orthologous genes (OGs) (Hachiya and Sakakibara, 2009). Our workflow was applied to large-scale comparisons of 101 prokaryotic and 15 fungal genomes. Thereafter, we examined the difference between OGs in conserved gene clusters (clustered OGs) and OGs that are not the members of conserved gene clusters (isolated OGs). Our analysis confirms the finding in previous studies that, in prokaryotes, protein sequences of clustered OGs are more conserved than those of isolated OGs. In addition, this interesting correlation between protein sequence homology and gene order conservation was observed also in fungal genomes. To our knowledge, this is the first report of a systematic survey of such correlation in eukaryotic genomes. Furthermore, we analyzed evolutionary forces behind the correlation by estimating the rate of synonymous substitutions (K_S) and the rate of nonsynonymous substitutions (K_A). This detailed sequence analysis reveals that although the correlation is consistently observed and seems to be a general trend among prokaryotic and fungal genomes, the evolutionary forces behind the correlation are different among lineages, suggesting that the joint effect of heterogeneous underlying mechanisms would result in the correlation.

1 Background

The rapid increase of the availability of completely sequenced genomes provides us with an opportunity to explore the underlying mechanisms for the evolution of genome organizations. Especially in prokaryotes, the number of completely sequenced genomes has been exponentially increased, with a doubling time of approximately 20 months for bacteria and approximately 34 months for archaea (Koonin and Wolf, 2008). As of this writing (9 May 2009), 812 bacterial and 58 archaeal genomes can be downloaded from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/genomes/). These collections of prokaryotic genomes cover 21 bacterial and four archaeal phyla, indicating that the current collections of bacterial and archaeal genomes provide a reasonable approximation of the diversity of prokaryotic life forms on earth (Koonin and Wolf, 2008).

Structural changes in complete genome sequences have been extensively examined, and it has been shown that large-scale gene orders (e.g. more than ten genes) are

hardly conserved even between closely related prokaryotic genomes (Mushegian and Koonin, 1996; Tatusov *et al.*, 1996; Watanabe *et al.*, 1997; Dandekar *et al.*, 1998; Koonin, 2009), suggesting that extensive gene shuffling has occurred during prokaryotic genome evolution (Koonin *et al.*, 1996). On the other hand, gene orders of a few neighboring genes have been preserved even between distantly related prokaryotic genomes, and physical interactions between the proteins encoded by genes in such conserved gene clusters are apparent in most cases (Dandekar *et al.*, 1998). Based on this observation, the information of the gene order conservation has been used to complement homology-based prediction of protein functions (Huynen *et al.*, 2000; Wolf *et al.*, 2001; Li *et al.*, 2007). Whereas the homology of protein sequences can be used to predict the molecular function of a protein, the gene order conservation can be used to predict a higher order function (e.g. in which process or pathway a particular protein plays a role, or with which other protein it interacts) (Huynen *et al.*, 2000). Deepening the understanding of the evolutionary forces that preserve gene orders would provide us with valuable biological insights, which can be used to increase the accuracy of the protein function prediction based on the gene order conservation.

Here, we are focusing on an interesting finding that links between the evolution of protein sequences and the evolution of gene orders. Dandekar *et al.* (1998) performed a systematic comparison of nine bacterial and archaeal genomes, and found that the degree of protein sequence conservation of genes in conserved gene clusters is on average substantially higher than that of the other genes. More recently, Lemoine *et al.* (2007) corroborated this finding by comparing 107 bacterial and archaeal genomes. This finding would be an important clue toward unraveling the evolutionary forces that preserve gene orders. However, the previous studies do not conduct further analyses for discussing evolutionary forces behind the correlation between protein sequence homology and gene order conservation.

In the present study, we shed light on the evolutionary forces by estimating the rate of synonymous substitutions (K_S) and the rate of nonsynonymous substitutions (K_A). The ratio between K_A and K_S (K_A/K_S) can be used to assess how strong evolutionary pressures have enforced conservation of protein sequences because $K_A/K_S = 1$ means neutral mutations, $K_A/K_S < 1$ purifying selections, and $K_A/K_S > 1$ diversifying positive selections (Yang *et al.*, 2000). We can also assess how frequently the coding sequence of a gene has been substituted based on the value of K_S. We here assume that higher degree of protein sequence conservation of clustered OGs can be explained by either stronger selective pressures to maintain protein sequences (lower value of K_A/K_S ratio), lower substitution rate of coding sequences (lower value of K_S), or both. Based upon this assumption, we aim at unraveling which of the three explanations is appropriate for each taxonomic group.

For this purpose, we propose a novel workflow which enables sensitive detection of conserved gene clusters. Our workflow uses the OASYS program in order to identify orthologous genes. OASYS can accurately detect one-to-one orthology relationships of genes by taking into account the information of gene order conservation. This makes it possible to avoid too stringent criteria for filtering out suspicious homologs,

and to detect conserved gene clusters sensitively. The source code of OASYS is freely available at http://oasys.dna.bio.keio.ac.jp under the GNU General Public License.

In addition, we included fungal genomes in our analyses, enabling us to discuss how general the finding in Dandekar *et al.* (1998) is in a wide variety of species, including not only prokaryotes but also eukaryotes. The correlation between protein sequence homology and gene order conservation in eukaryotes has been less intensively surveyed than in prokaryotes. Hillier *et al.* (2007) reports a slightly related finding that the sequence conservation rate of syntenic OGs is higher than that of non-syntenic OGs in the comparison of nematodes, where syntenic OGs are defined as the OGs located on the corresponding chromosomes of different species. Note that our definition of clustered OGs and their definition of syntenic OGs are substantially different. To our knowledge, our analyses of fungi genomes are the first attempt to survey in a systematic manner whether the finding in Dandekar *et al.* (1998) can be extended to eukaryotes.

2 OASYS: Orthology Assignment based on Synteny and Sequence Information

Bandyopadhyay *et al.* (2006) identifies *functional orthologs*, which are genes in different species that play functionally equivalent roles, on the basis of the concept that a protein and its functional ortholog are likely to interact with proteins in their respective networks that are themselves functional orthologs. Analogously, OASYS identifies *positional orthologs*, which are genes in different species that are located on corresponding chromosomal positions, on the basis of the concept that a gene and its positional ortholog are likely to be located on their respective chromosomal positions that are diagonally proximate to themselves positional orthologs.

In order to identify positional orthologs based not only on the information of protein sequence conservation but also on the information of gene order conservation, we propose a novel algorithm named OASYS (Orthology Assignment based on SYnteny and Sequence information). The OASYS algorithm executes the following procedures:

1. Detect homologous gene pairs by comparing all protein sequences encoded by a genome G_A and all protein sequences encoded by a genome G_B. With the default setting, gene pairs whose bit score is greater than 50 bits are detected as homologous gene pairs.
2. Detect seed orthologs by applying the reciprocal best BLAST hit (RBH) method.
3. Quantify the extent of gene order conservation by computing the *weighted number of neighboring seed orthologs* (WNNSO) value for each homologous gene pair. The method to compute WNNSO values is described in the section 2.1.
4. Estimate probability densities of the bit scores and the WNNSO values. The probability density functions used in OASYS are described in the section 2.2. A method to fit the observed bit scores or WNNSO values to these probability

density functions is described in the section 2.3.

5. Calculate the integrated conservation score, which takes into account the gene order conservation as well as the protein sequence conservation, for each homologous gene pair. The scoring scheme used in OASYS is described in the section 2.4.
6. Detect reciprocal best similarity pairs in terms of integrated conservation scores as positional orthologs.

2.1 Weighted number of neighboring seed orthologs

OASYS quantifies the extent of gene order conservation by a novel measure named weighted number of neighboring seed orthologs (WNNSO). Given a set of homologous gene pairs and its bit scores, the calculation of the WNNSO values starts with the detection of putative orthologs. Putative orthologs are simply detected by the reciprocal best hit (RBH) method, that is, reciprocal best similarity pairs in terms of bit scores are detected as putative orthologs (Rivera *et al.*, 1998; Hirsh and Fraser, 2001; Jordan *et al.*, 2002). We call the putative orthologs '*seed orthologs*', and the homologous gene pairs that are not identified as putative orthologs '*non-seed homologs*'.

Second, the diagonal proximity between homologous gene pairs and the seed orthologs are computed on the basis of the matrix representation of gene positions. Let A be a set of genes encoded by the genome G_A, A^k be a set of genes located on the k-th chromosome of the genome G_A, and a_i^k be the i-th gene located on the k-th chromosome. We assume without loss of generality that the elements in A^k are sorted in order of increasing start position along the k-th chromosome. Regarding genome G_B, B, B^l, and b_j^l are similarly defined. Then, a homologous gene pair (a_i^k, b_j^l) is represented as an element of a $|A^k| \times |B^l|$ matrix, in which a homologous gene pair (a_i^k, b_j^l) corresponds to a point (i, j). If two gene pairs $h_m = (a_i^k, b_j^l)$ and $h_{m'} = (a_{i'}^{k'}, b_{j'}^{l'})$ are *collinear*, a special distance function named diagonal pseudo distance (DPD) (Vandepoele *et al.*, 2002) is used to define the distance between the two gene pairs:

$$\mathrm{DPD}(h_m, h_{m'}) = 2\max(|i - i'|, |j - j'|) - \min(|i - i'|, |j - j'|). \tag{III.1}$$

If two gene pairs are not collinear, the distance is defined as infinity. The definition of the 'collinearity' can be found in the section 2.1.1.

Finally, the WNNSO value is computed for each homologous gene pair by counting the number of the seed orthologs near the homologous gene pair with weights that decrease with increasing the diagonal pseudo distance. Let S be a set of seed orthologs. Then, the WNNSO value for a homologous gene pair h_m is given by

$$\mathrm{WNNSO}(h_m|S) = \sum_{h_{m'} \in S} \mathrm{Weight}(h_m, h_{m'}) \tag{III.2}$$

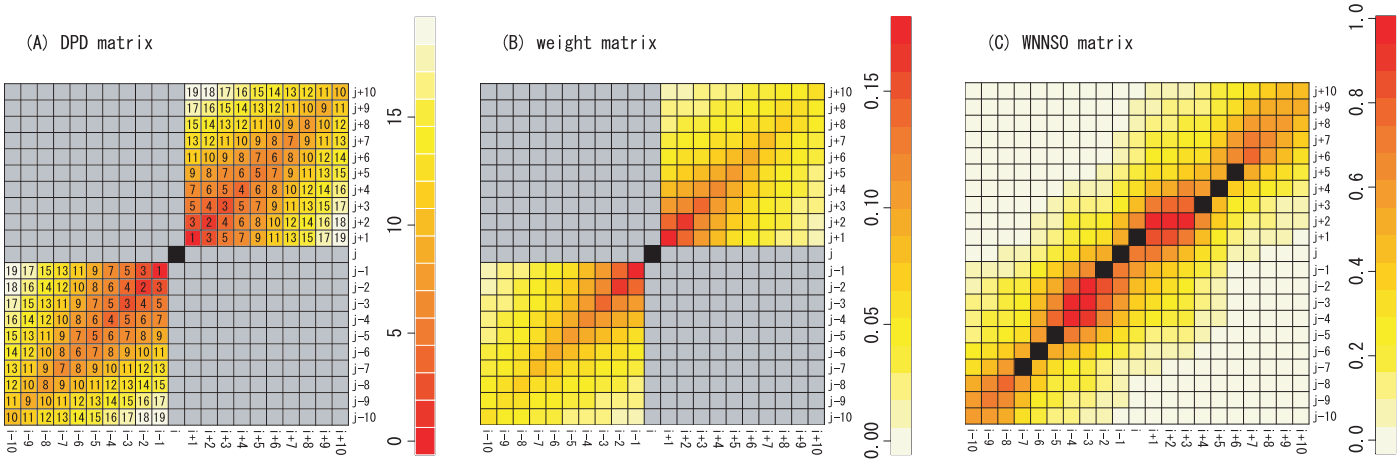


Fig. III.1 Graphical representation of the DPD, weight and WNNSO functions. (A) DPD functions. Given a homologous gene pair $h_m = (x_i, y_j)$ (represented as the central element colored by black), the color of an element $h_{m'} = (i', j')$ in the matrix represents the degree of the value of $DPD(h_m, h_{m'})$. Here, we assume that $h_m.sign = 1$ and $h_{m'}.sign = 1$. Positive integer shown in each element is the DPD value. Gray-colored elements in the matrix correspond to the gene pairs that are not collinear with the gene pair (x_i, y_j). The DPD value in these elements is defined as infinity. (B) Weight function. The color of an element $h_{m'} = (i', j')$ in the matrix represents the degree of the value of Weight$(h_m, h_{m'})$. When computing the weight values, the value of σ parameter was set at 2.0, and the value of Cut_dpd parameter was set at 20. Regarding gray-colored elements, the value of Weight$(h_m, h_{m'})$ was computed as zero. (C) WNNSO function. Given a set of seed orthologs S (represented by elements colored by black), the color of each element $h_{m'}$ represents the degree of the value of WNNSO$(h_{m'}|S)$.

with

$$\text{Weight}(h_m, h_{m'}) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{\text{DPD}(h_m, h_{m'})}{2\sigma^2}) & \text{when DPD}(h_m, h_{m'}) \leq \text{Cut\_dpd} \\ 0 & \text{otherwise ,} \end{cases} \quad \text{(III.3)}$$

where σ and Cut_dpd are user-defined parameters. σ controls the degree of the decrease of the weight value with increasing DPD, and Cut_dpd represents the threshold for DPD values. Note that the weight between non-collinear gene pairs becomes zero. Fig. III.1 shows the result of applying the above DPD, weight and WNNSO functions on a hypothetical example.

2.1.1 Collinearity

If a gene x_i^k is located on the forward strand of the k-th chromosome, we denote $x_i^k.strand = 1$. If x_i^k is located on the reverse strand, we denote $x_i^k.strand = -1$. For a gene y_j^l, $y_j^l.strand$ is similarly defined. Then, the *sign* of a gene pair $h_m = (x_i, y_j)$

is defined as $h_m.sign = x_i.strand \times y_j.strand$. OASYS defines that two gene pairs $h_m = (x_i^k, y_j^l)$ and $h_{m'} = (x_{i'}^{k'}, y_{j'}^{l'})$ are collinear if the following conditions are satisfied:

$$\begin{aligned} & k = k', \quad l = l', \quad i \neq i', \quad j \neq j', \\ & h_m.sign = h_{m'}.sign, \quad \frac{j - j'}{i - i'} \times h_m.sign > 0. \end{aligned} \tag{III.4}$$

2.1.2 Effect of σ parameter

Fig. III.2 shows the result of applying the weight function on a hypothetical example with varying the value of the σ parameter, in which we can observe the effect of the σ parameter on the weight function. When the σ parameter takes a small value (e.g. <= 1.0), the weight function decreases with increasing DPD value so rapidly that the weight value becomes almost zero even in the points that are not so distant from the center point. In other words, in the computation of the WNNSO value at the point (i, j), only a few seed orthologs, which are located on very near the point (i, j), is counted, and the information on the other seed orthologs will be discarded. The loss of such information would make the WNNSO value less robust measure of the gene order conservation to the method to detect seed orthologs. For example, in the case where the genuine positional orthologs very near the point (i, j) is not detected as seed orthologs, the WNNSO value at the point (i, j) would largely affected by the miss of the method to detect seed orthologs.

On the other hand, when the σ parameter takes a large value (e.g. >= 4.0), weight values are almost the same among all points shown in Fig. III.2 regardless of the DPD value. In other words, seed orthologs that are located on collinear positions are equally counted in the computation of the WNNSO value. The loss of information on the diagonal proximity would make the WNNSO value less valuable measure of the gene order conservation.

When σ parameter takes the value between 2.0 and 3.0, both information can be taken into account; the weight value gradually decreases with increasing DPD value so that the weight of closer points is larger than the weight of more distant points, and the weight is not almost zero even in relatively distant points. Thus, the default value of the σ is set at 2.0 in the current version of OASYS.

Within such appropriate range, the σ parameter controls the trade-off between the importance of two sets of information: a small value of the σ parameter enhances the degree of the decrease of weight value with increasing DPD value. Thus, the information of seed orthologs located on very near positions is counted with a larger weight. When comparing two genomes that are closely related, a small value of σ parameter is expected to decrease the WNNSO value of gene pairs that are not genuine positional ortholog. A large value of σ parameter weakens the degree of the decrease of weight value with increasing DPD value. Thus, the information of seed orthologs located on distant positions to somewhat is also counted. Thus, when comparing two genomes that are distantly related, a large value of σ parameter is expected to increase the sensitivity of the collinearity.

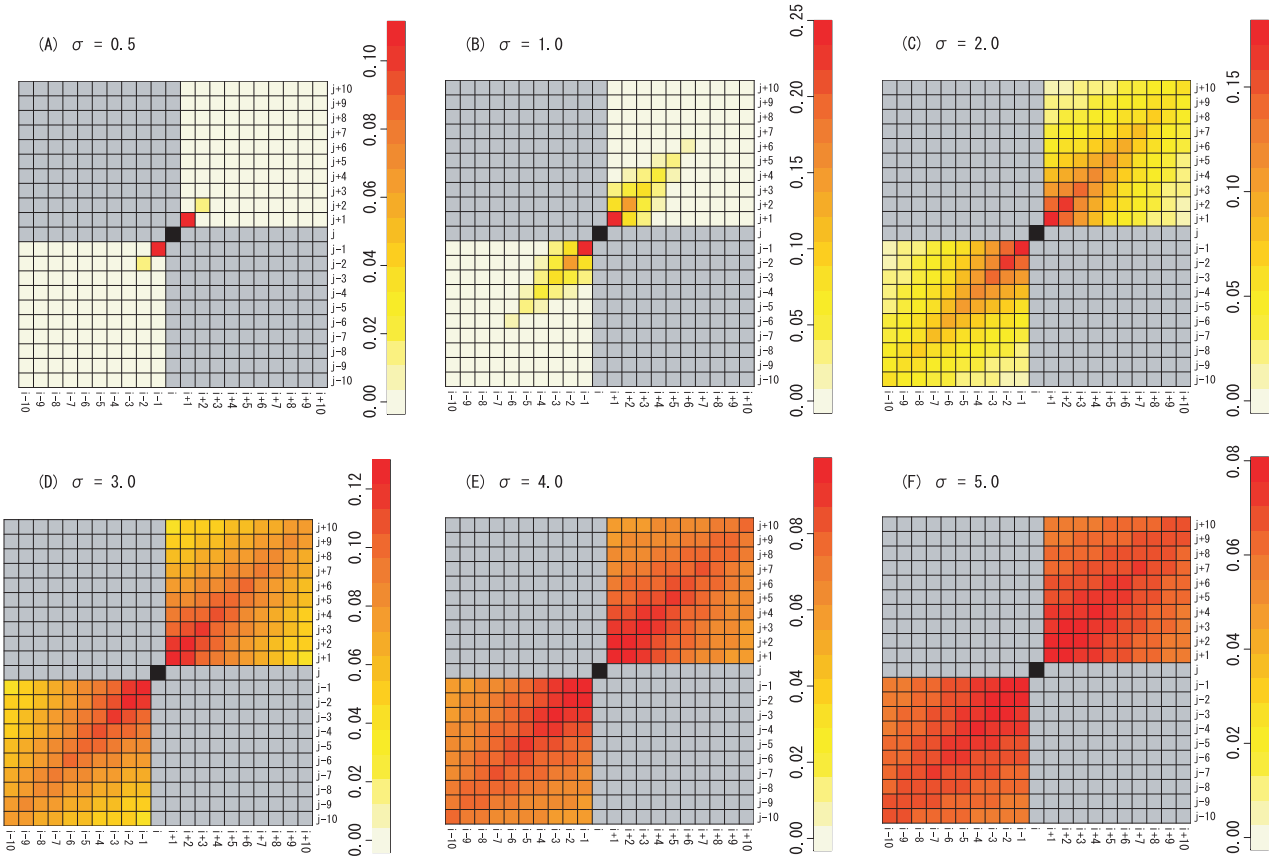


Fig. III.2 Effect of the σ parameter on the weight function. These figures show the matrices in which the color of each point (i', j') represents the degree of the value of the weight function $W(h_{m'}|h_m)$, where $h_m = (x_i, y_j)$ corresponds to the point (i, j), which is represented as the central point colored in black, and $h_{m'} = (x_{i'}, y_{j'})$ corresponds to each point (i', j'). Here, h_m and $h_{m'}$ are assumed to have a positive orientation ($h_m.sign = 1$ and $h_{m'}.sign = 1$). The points colored in gray are not collinear points with the central point (i, j), and therefore, the weight value in these points becomes zero regardless of the value of the σ parameter. The computation of the weight function was performed with varying the value of the σ parameter: (A) $\sigma = 0.5$, (B) $\sigma = 1.0$, (C) $\sigma = 2.0$, (D) $\sigma = 3.0$, (E) $\sigma = 4.0$, and (F) $\sigma = 5.0$.

2.2 Probability density functions

In order to distinguish between positional orthologs and other homologs, OASYS takes advantage of the difference in the extent of the gene order conservation between positional orthologs and other homologs. For this end, OASYS assumes that the probability density of the WNNSO values for genuine positional orthologs can be approximated by that for seed orthologs. It is also assumed that the probability density of the WNNSO values for other homologs can be approximated by that for non-seed homologs.

In addition to the difference in the extent of gene order conservation, OASYS also makes use of the difference in the extent of protein sequence conservation between positional orthologs and other homologs. As in the case of the WNNSO values, OASYS assumes that the probability density of the bit scores for positional orthologs (for other homologs) can be approximated by that for seed orthologs (for non-seed homologs).

Accordingly, OASYS models four probability densities; (i) the probability density of the WNNSO values for positional orthologs, (ii) the probability density of the bit scores for positional orthologs, (iii) the probability density of the WNNSO values for other homologs, and (iv) the probability density of the bit scores for other homologs. Each of the four probability densities is modeled by either of two probability density functions (pdfs), namely the *one-sided generalized Gaussian* (OGG) pdf and the *asymmetric generalized Gaussian* (AGG) pdf. As shown later, the former can represent wide range of decreasing functions, and the later can represent wide range of unimodal functions. The choice of the model is performed on the basis of the Akaike information criteria (Akaike, 1974). Figs. III.3 and III.4 show that our model is well fitted to each data set.

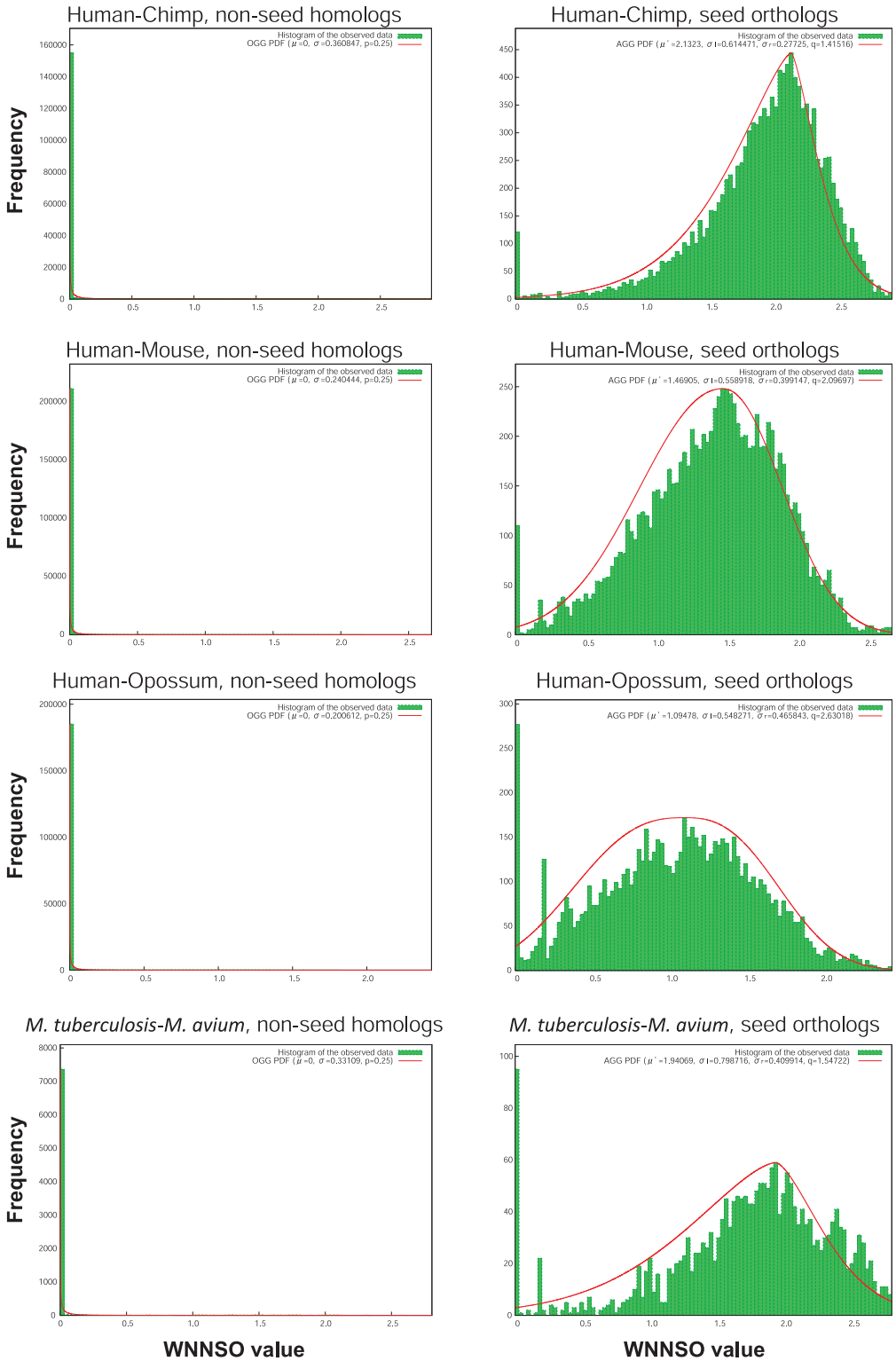


Fig. III.3 Histograms of WNNSO values together with probability density functions. Seed orthologs were detected by the traditional RBH method. Non-seed homologs were defined as the homologous gene pairs that are not seed orthologs. WNNSO values were calculated by setting the σ parameter at 2.0 and the CUT_DPD parameter at 10. The model selection and the parameter optimization are performed based on the Akaike information criteria and the conjugate gradient method, respectively.

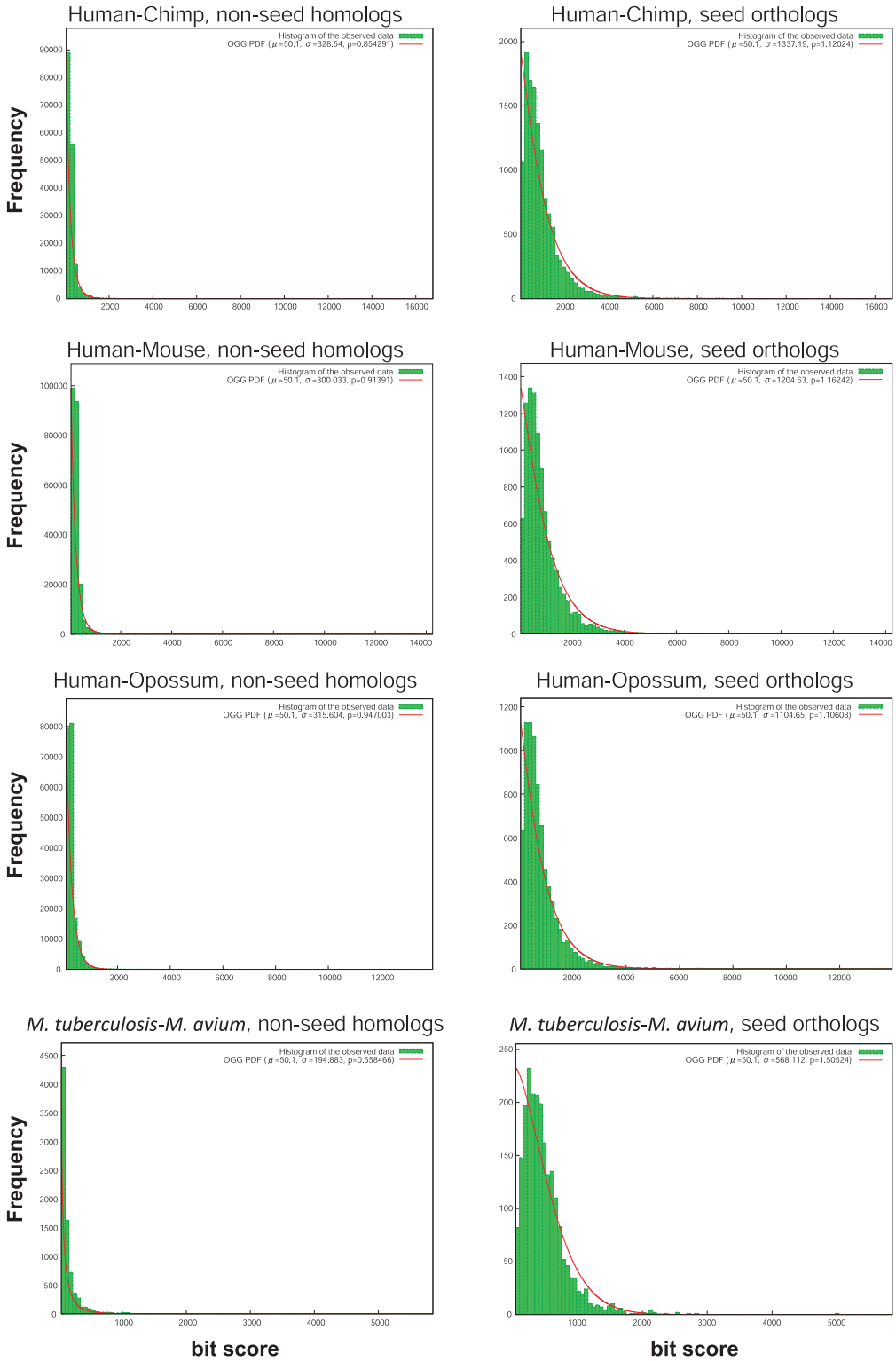


Fig. III.4 Histograms of bit scores together with probability density functions. Seed orthologs were detected by the traditional RBH method. Non-seed homologs were defined as the homologous gene pairs that are not seed orthologs. Bit scores were calculated by the BLASTP program. The model selection and the parameter optimization are performed based on the Akaike information criteria and the conjugate gradient method, respectively.

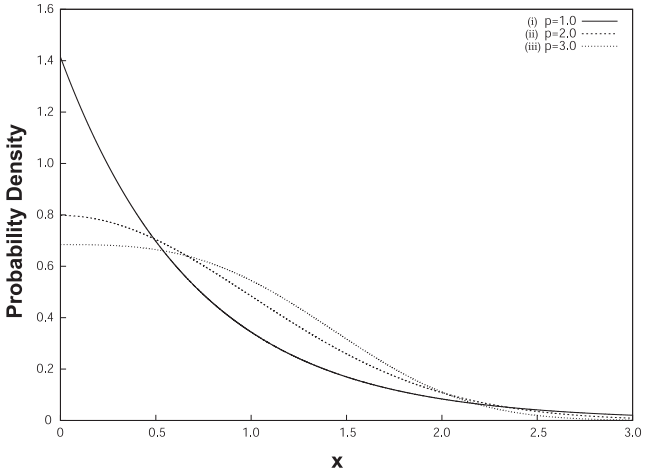


Fig. III.5 One-sided generalized Gaussian pdf. Three curves shown in this figure have the following parameter values; (i) $\mu = 0$, $\sigma^2 = 1$, and $p = 1.0$, (ii) $\mu = 0$, $\sigma^2 = 1$, and $p = 2.0$, (iii) $\mu = 0$, $\sigma^2 = 1$, and $p = 3.0$.

2.2.1 One-sided generalized Gaussian distribution

The generalized Gaussian (GG) distribution proposed in Miller and Thomas (1972) is given by

$$P_{\rm gg}(x;\mu,\sigma,p) = \begin{cases} \frac{p\gamma}{2\Gamma(\frac{1}{p})}\exp(-\gamma^p(\mu-x)^p) & \text{when } x<\mu \\ \frac{p\gamma}{2\Gamma(\frac{1}{p})}\exp(-\gamma^p(x-\mu)^p) & \text{when } x\geq\mu, \end{cases} \tag{III.5}$$

where $\gamma = \frac{1}{\sigma}\sqrt{\frac{\Gamma(\frac{3}{p})}{\Gamma(\frac{1}{p})}}$ and $\Gamma(\bullet)$ is the gamma function. In this model, μ, σ^2, and p denote the mean, variance, and decay rate (also referred to as shape parameter) of the pdf, respectively. We modify Eq. (III.5) and define the one-sided generalized Gaussian (OGG) distribution, which is given by

$$P_{\rm ogg}(x;\mu,\sigma,p) = \begin{cases} 0 & \text{when } x<\mu \\ \frac{p\gamma}{\Gamma(\frac{1}{p})}\exp(-\gamma^p(x-\mu)^p) & \text{when } x\geq\mu. \end{cases} \tag{III.6}$$

Note that μ in Eq. (III.6) is not the mean of the OGG pdf but a location parameter. For $x \geq \mu$, the pdf is a decreasing function of x. As shown in Fig. III.5, the OGG family of distributions can represent wide range of decreasing functions by changing the shape parameter p.

Suppose that we are given an observed data set of scalar values $x = \{x_1, \ldots, x_N\}$ and that $x_i \geq \mu$ for $1 \leq i \leq N$. Then, the log likelihood function of the OGG pdf is given by

$$\ln L_{\rm ogg} = N \ln\left(\frac{p\gamma}{\Gamma(\frac{1}{p})}\right) - \sum_{i=1}^{N}\gamma^p(x_i-\mu)^p. \tag{III.7}$$

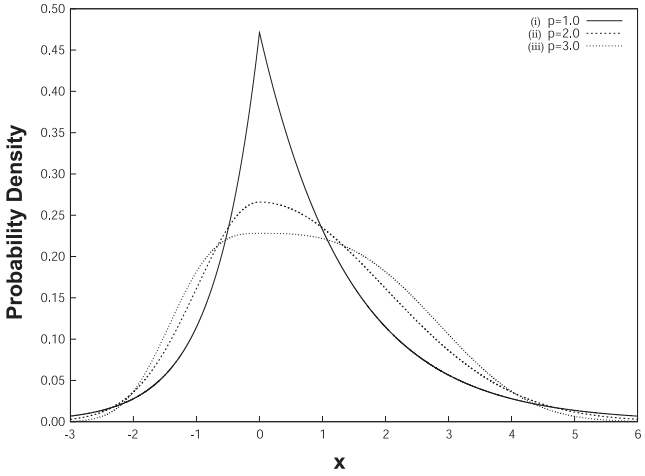


Fig. III.6 Asymmetric generalized Gaussian pdf. Three curves shown in this figure have the following parameter values; (i) $\mu = 0$, $\sigma_l^2 = 1$, $\sigma_r^2 = 2$ and $p = 1.0$, (ii) $\mu = 0$, $\sigma_l^2 = 1$, $\sigma_r^2 = 2$ and $p = 2.0$, (iii) $\mu = 0$, $\sigma_l^2 = 1$, $\sigma_r^2 = 2$ and $p = 3.0$.

We can optimize the parameters in the OGG model so as to maximize Eq. (III.7). For details, see the section 2.3.

2.2.2 Asymmetric generalized Gaussian distribution

The asymmetric generalized Gaussian (AGG) distribution proposed in Tesei and Regazzoni (1998) is given by

$$P_{\text{agg}}(x; \mu', \sigma_l, \sigma_r, q) = \begin{cases} \frac{q\gamma_a}{\Gamma(\frac{1}{q})} \exp(-\gamma_l^q(-x+\mu')^q) & \text{when } x < \mu' \\ \frac{q\gamma_a}{\Gamma(\frac{1}{q})} \exp(-\gamma_r^q(x-\mu')^q) & \text{when } x \geq \mu', \end{cases} \tag{III.8}$$

where $\gamma_a = \frac{1}{\sigma_l+\sigma_r}\sqrt{\frac{\Gamma(\frac{3}{q})}{\Gamma(\frac{1}{q})}}$, $\gamma_l = \frac{1}{\sigma_l}\sqrt{\frac{\Gamma(\frac{3}{q})}{\Gamma(\frac{1}{q})}}$, and $\gamma_r = \frac{1}{\sigma_r}\sqrt{\frac{\Gamma(\frac{3}{q})}{\Gamma(\frac{1}{q})}}$. In this model, μ' is the mode, σ_l^2 and σ_r^2 are the variances of the left and right side respectively, and q is the decay rate. It is noticed that if $\sigma_l^2 = \sigma_r^2$ then the pdf coincides with the GG distribution, hence it is symmetric (Lee and Nandi, 1999). For the symmetric cases, $c = 2$ represents the Gaussian distribution while $c = 1$ represents the Laplace distribution. If $\sigma_l^2 \neq \sigma_r^2$ then the pdf represents an asymmetric model. As shown in Fig. III.6, the AGG family of distributions can represent wide range of unimodal probability density functions by changing the shape parameter q.

Suppose that we are given an observed data set of scalar values $x = \{x_1, \ldots, x_N\}$.

Then, the log likelihood function of the AGG pdf is given by

$$\begin{aligned}\ln L_{\mathrm{agg}} = N \ln\left(\frac{q\gamma_a}{\Gamma(\frac{1}{q})}\right) &- \sum_{i=1, x_i<\mu'}^{N} \gamma_l^q (\mu' - x_i)^q \\ &- \sum_{i=1, x_i\geq\mu'}^{N} \gamma_r^q (x_i - \mu')^q.\end{aligned} \tag{III.9}$$

We can optimize the parameters in the AGG model so as to maximize Eq. (III.9). For details, see the section 2.3.

2.3 Parameter optimization and model selection

Suppose that we are given an observed data set of scalar values $x = \{x_1, \ldots, x_N\}$ and we want to fit the observed data to the one-sided generalized Gaussian (OGG) model or the asymmetric generalized Gaussian (AGG) model. Furthermore, suppose that we want to know which model, either of the OGG model or the AGG model, is more appropriate to represent the probability density of the observed data.

2.3.1 Fitting to an OGG distribution

Parameters in the OGG model can be optimized by maximizing the log likelihood function shown in Eq. (III.7). At first, the location parameter μ is naturally defined as the minimum value of the observed data because the OGG pdf is one-sided distribution. Secondly, the variance σ^2 is defined as the second-order moment of the observed scalar values around the location parameter μ. Thus, σ^2 is given by

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{III.10}$$

Finally, the shape parameter p is optimized by using a conjugate gradient method. The derivative of Eq. (III.7) with respect to the shape parameter p is given by

$$\frac{\partial}{\partial p}\ln L_{\mathrm{ogg}} = D_1 - \sum_{i=1}^{N}\gamma^p (x_i - \mu)^p [\ln\{\gamma(x_i - \mu)\} + D_2] \tag{III.11}$$

with

$$\begin{aligned}D_1 &= \frac{1}{2}N\left(\frac{2}{p} - 3\frac{\Psi(\frac{3}{p})}{p^2} + 3\frac{\Psi(\frac{1}{p})}{p^2}\right) \\ D_2 &= \frac{\Psi(\frac{1}{p}) - 3\Psi(\frac{3}{p})}{2p},\end{aligned}$$

where $\Psi(\bullet)$ is the digamma function and defined as $\Psi(\bullet) = \frac{\partial \ln\Gamma(\bullet)}{\partial(\bullet)} = \frac{\partial\Gamma(\bullet)}{\Gamma(\bullet)\partial(\bullet)}$. Conjugate gradient methods can find a local optimum solution with the knowledge of

Eq. (III.7) and Eq. (III.11). In our implementation of the OASYS algorithm, the Fletcher-Reeves conjugate gradient algorithm (Fletcher and Reeves, 1964), which is one of the most basic algorithm for non-quadratic optimization, is used to optimize the shape parameter p.

2.3.2 Fitting to an AGG distribution

Parameters in the AGG model can be optimized by maximizing the log likelihood function shown in Eq. (III.9). At first, the mode μ' is estimated by using the histogram method, which estimates the mode to be the value of the bin with the greatest number of data points (Hedges and Shah, 2003). Secondly, the variance parameters σ_l^2 and σ_r^2 are calculated by

$$\sigma_l^2 = \frac{1}{N_l - 1} \sum_{i=1, x_i < \mu'}^{N} (x_i - \mu')^2 \tag{III.12}$$

$$\sigma_r^2 = \frac{1}{N_r - 1} \sum_{i=1, x_i \geq \mu'}^{N} (x_i - \mu')^2, \tag{III.13}$$

where N_l (or N_r) is the number of the observed data points having the values smaller (or greater) than the mode parameter μ'. Thus, N_l and N_r are given by

$$N_l = \sum_{i=1, x_i < \mu'}^{N} 1 \tag{III.14}$$

$$N_r = \sum_{i=1, x_i \geq \mu'}^{N} 1. \tag{III.15}$$

Finally, the shape parameter q is optimized by using a conjugate gradient method. As shown in (Lee and Nandi, 1999), the derivative of Eq (III.9) with respect to the shape parameter q is given by

$$\begin{aligned} \frac{\partial}{\partial q} \ln L_{\text{agg}} = D_1' &- \sum_{i=1, x_i < \mu'}^{N} \gamma_l^q (\mu' - x_i)^q [\ln\{\gamma_l(\mu' - x_i)\} + D_2'] \\ &- \sum_{i=1, x_i \geq \mu'}^{N} \gamma_r^q (x_i - \mu')^q [\ln\{\gamma_r(x_i - \mu')\} + D_2'] \end{aligned} \tag{III.16}$$

with

$$D_1' = \frac{1}{2} N \left(\frac{2}{q} - 3\frac{\Psi(\frac{3}{q})}{q^2} + 3\frac{\Psi(\frac{1}{q})}{q^2} \right)$$

$$D_2' = \frac{\Psi(\frac{1}{q}) - 3\Psi(\frac{3}{q})}{2q}.$$

By using Eq. (III.9), Eq. (III.16), and the Fletcher-Reeves conjugate gradient algorithm, our implementation of the OASYS algorithm finds a local optimum solution for the shape parameter q.

2.3.3 Model selection

Let $\tilde{\boldsymbol{\theta}}_{\mathrm{ogg}}$ be the optimal parameter set for the OGG model and $\tilde{\boldsymbol{\theta}}_{\mathrm{agg}}$ be the optimal parameter set for the AGG model. OASYS determines which model, either of the OGG model or the AGG model, is more appropriate for representing the probability density of the observed data based upon the Akaike information criteria (Akaike, 1974). If the following inequality is satisfied, the OGG model is selected:

$$\ln P_{\mathrm{ogg}}(x|\tilde{\boldsymbol{\theta}}_{\mathrm{ogg}}) - M_{\mathrm{ogg}} > \ln P_{\mathrm{agg}}(x|\tilde{\boldsymbol{\theta}}_{\mathrm{agg}}) - M_{\mathrm{agg}}, \tag{III.17}$$

where M_{ogg} and M_{agg} are the number of adjustable parameters in the OGG model and the AGG model, respectively. Thus, $M_{\mathrm{ogg}} = 3$ and $M_{\mathrm{agg}} = 4$. Note that $\ln P_{\mathrm{ogg}}(x|\tilde{\boldsymbol{\theta}}_{\mathrm{ogg}})$ and $\ln P_{\mathrm{agg}}(x|\tilde{\boldsymbol{\theta}}_{\mathrm{agg}})$ can be evaluated based on Eqs. (III.7) and (III.9), respectively. If the inequality shown in Eq. (III.17) is not satisfied, the AGG model is selected.

2.4 Scoring scheme

Given the model for describing the probability density of the WNNSO values of positional orthologs M^{+}_{wnnso} and the model for describing the probability density of the WNNSO values of other homologs M^{-}_{wnnso}, the *synteny score* of a homologous gene pair h_m whose WNNSO value is x is defined by

$$\mathrm{Syn\_Score}(h_m) = \ln \frac{P(x|M^{+}_{\mathrm{wnnso}})}{P(x|M^{-}_{\mathrm{wnnso}})}. \tag{III.18}$$

As shown in Fig. III.7, the score function given by Eq. (III.18) is not necessary monotonically increasing function of x, although the score function is desired to be monotonically increasing function because it is considered that the homologous gene pairs which have greater WNNSO value are more likely to be positional orthologs. Thus, we modify Eq. (III.18) so that the score function be monotonically increasing. The modified score function is given by

$$\begin{aligned} &\mathrm{Modified\_Syn\_Score}(h_m) \\ &= \begin{cases} \ln \frac{P(x|M^{+}_{\mathrm{wnnso}})}{P(x|M^{-}_{\mathrm{wnnso}})} & \text{for } x \le \hat{x} \\ \ln \frac{P(\hat{x}|M^{+}_{\mathrm{wnnso}})}{P(\hat{x}|M^{-}_{\mathrm{wnnso}})} + \delta(x - \hat{x}) & \text{for } x > \hat{x}, \end{cases} \end{aligned} \tag{III.19}$$

where $\hat{x}$ is the WNNSO value at which the score function given by Eq. (III.18) takes the maximum value, and δ is a extremely small value. Fig. III.7 demonstrates that the

modified score function given by Eq. (III.19) is a monotonically increasing function of x.

Analogously, given the model for describing the probability density of the bit scores of positional orthologs M^{+}_{bit} and the model for describing the probability density of the bit scores of other homologs M^{-}_{bit}, the *sequence score* of a homologous gene pair h_m whose bit score is x is defined by

$$\textsc{Seq\_Score}(h_m) = \ln \frac{P(x|M^{+}_{\text{bit}})}{P(x|M^{-}_{\text{bit}})}. \tag{III.20}$$

As shown in Fig. III.8, the score function given by Eq. (III.20) is not necessary monotonically increasing function of x. Thus, we modify Eq. (III.20) so that the score function be monotonically increasing. The modified score function is given by

$$\textsc{Modified\_Seq\_Score}(h_m) = \begin{cases} \ln \frac{P(x|M^{+}_{\text{bit}})}{P(x|M^{-}_{\text{bit}})} & \text{for } x \leq \hat{x} \\ \\ \ln \frac{P(\hat{x}|M^{+}_{\text{bit}})}{P(\hat{x}|M^{-}_{\text{bit}})} + \delta(x - \hat{x}) & \text{for } x > \hat{x} \end{cases} \tag{III.21}$$

where $\hat{x}$ denote the bit score at which the score function given by Eq. (III.20) takes maximum value, and δ is a extremely small value. Fig.III.8 demonstrates that the modified score function given by Eq. (III.21) is a monotonically increasing function of x.

OASYS integrates the information about the extent of the gene order conservation and the extent of the protein sequence conservation by taking the weighted sum of the modified synteny score given by Eq. (III.19) and the modified sequence score given by Eq. (III.21). The integrated score is given by

$$\begin{aligned} &\text{Integrated\_Score}(h_m) \\ &= w_{\text{syn}}\text{Modified\_Syn\_Score}(h_m) + w_{\text{seq}}\text{Modified\_Seq\_Score}(h_m), \end{aligned} \tag{III.22}$$

where w_{syn} and w_{seq} denote the weight for the modified synteny score and the modified sequence score, respectively. The OASYS program has the weight ratio option, which can specify the weight ratio $\frac{w_{\text{syn}}}{w_{\text{seq}}}$. The default value for the weight ratio is set at 1.0.

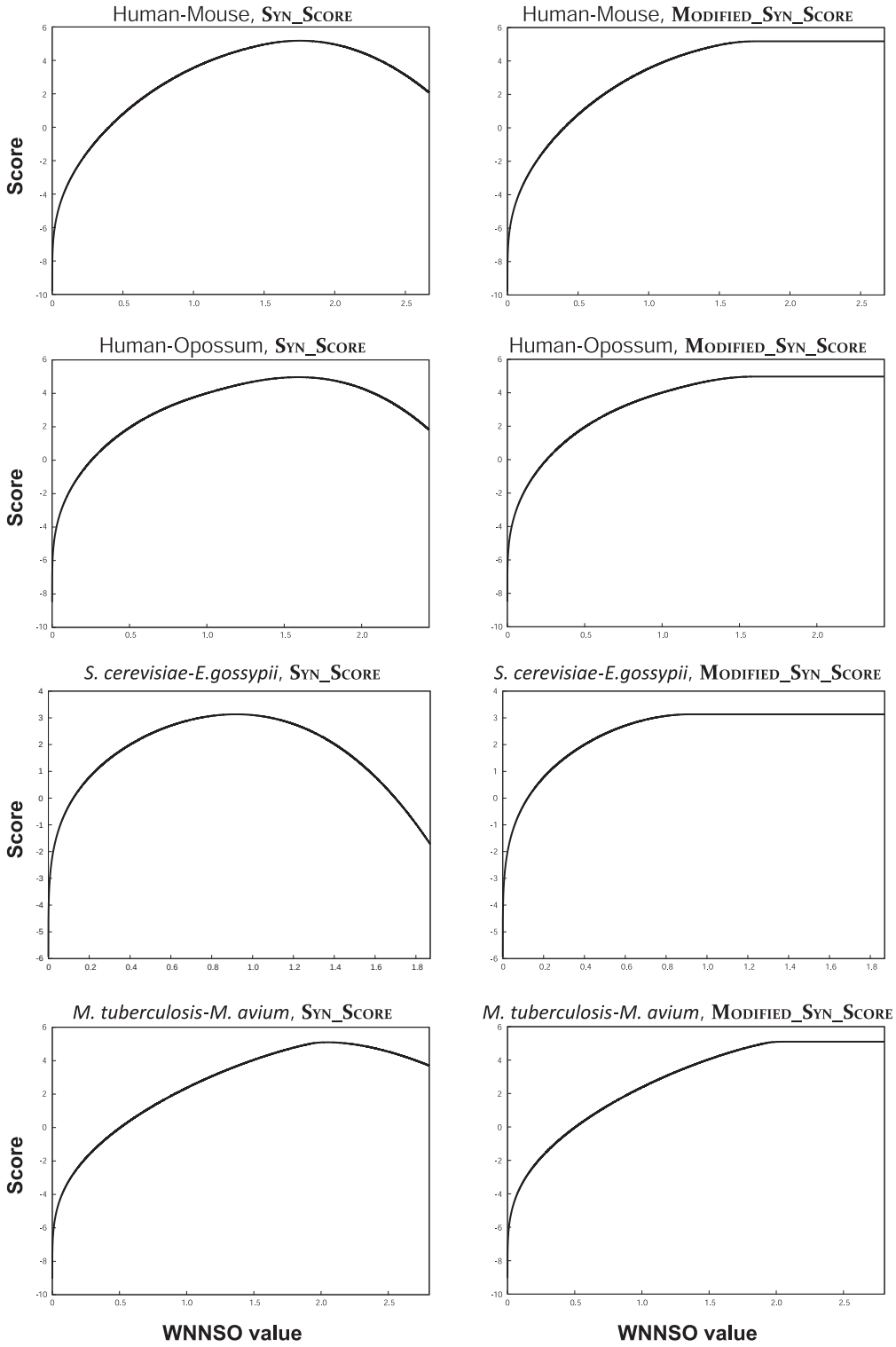


Fig. III.7 Plots of the synteny score function Syn_Score and the modified synteny score function Modified_Syn_Score. In these plots, horizontal axis shows the WNNSO value and vertical axis shows the synteny score or modified synteny score.

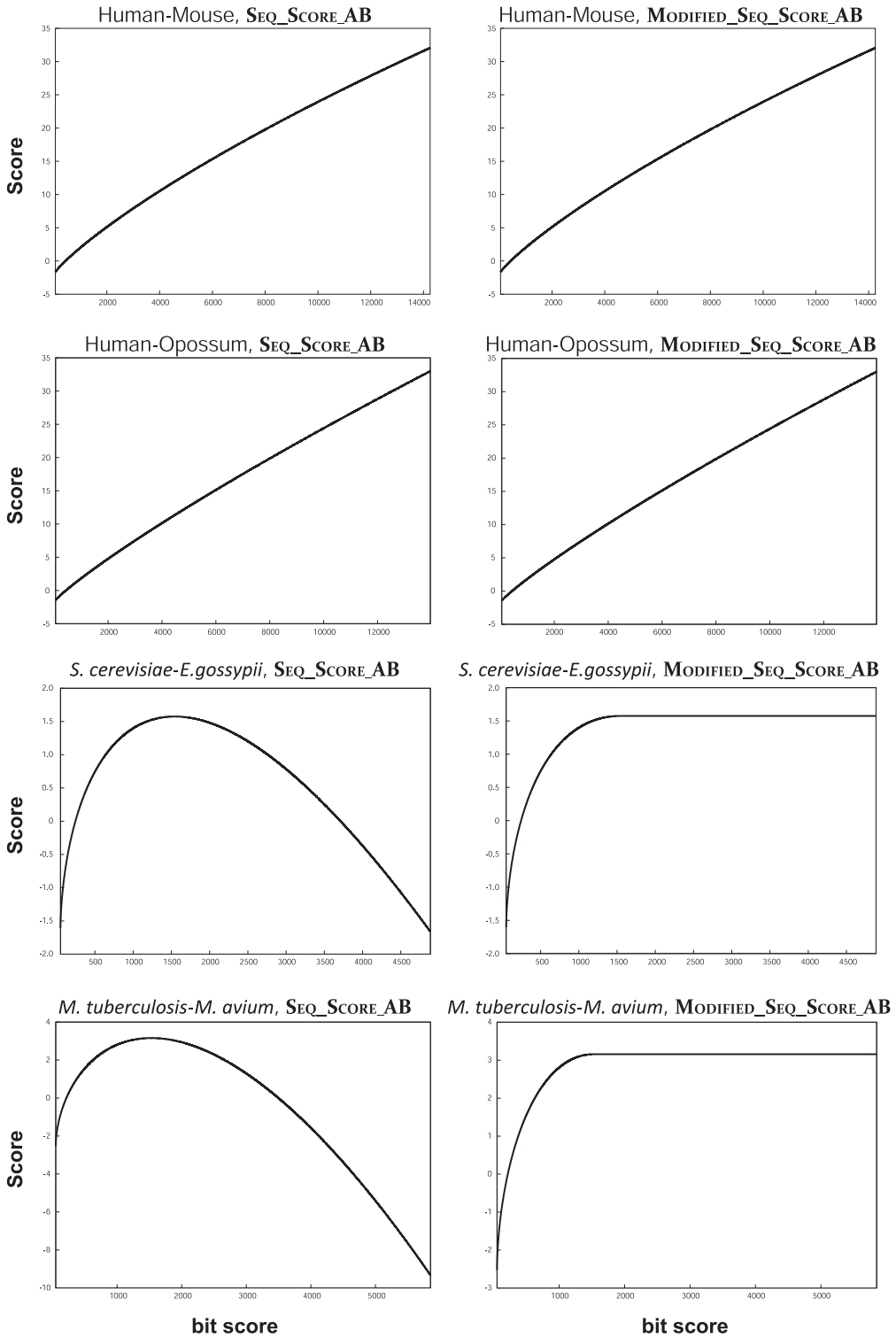


Fig. III.8 Plots of the sequence score function SEQ_SCORE, and the modified sequence score function MODIFIED_SEQ_SCORE. In these plots, horizontal axis shows the bit score and vertical axis shows the sequence score or modified sequence score.

3 Materials and Methods

3.1 Data resources

We downloaded complete sequences of bacterial, archaeal, and fungal genomes in GenBank format from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/genomes/). The taxonomic classification of these genomes was taken from the NCBI Taxonomy Browser (http://www.ncbi.nlm.nih.gov/Taxonomy/).

3.1.1 Bacterial genomes

Of 812 currently available bacterial genomes, 83 bacterial genomes were collected so as to cover all available bacterial orders (79 bacterial orders). These genomes cover 21 bacterial phyla, including two recently proposed bacterial phyla, Gemmatimonadetes (Zhang *et al.*, 2003) and Elusimicrobia (Herlemann *et al.*, 2009). A list of bacterial genomes used in our analyses and its taxonomy are shown in Table III.1.

Table. III.1 List of the bacterial genomes used in our analyses

Phylum	Class	Order	Species	Genome Size (kb)	# of Genes
Chloroflexi	Chloroflexi	Chloroflexales	*Chloroflexus aggregans*	4,685	3,730
		Herpetosiphonales	*Herpetosiphon aurantiacus*	6,785	5,278
	Dehalococcoidetes	not defined	*Dehalococcoides ethenogenes*	1,470	1,580
	Thermomicrobia	Thermomicrobiales	*Thermomicrobium roseum*	2,921	2,854
Deinococcus -Thermus	Deinococci	Deinococcales	*Deinococcus radiodurans*	3,284	3,167
		Thermales	*Thermus thermophilus*	2,116	2,238
Cyanobacteria	Gloeobacteria	Gloeobacterales	*Gloeobacter violaceus*	4,659	4,430
	not defined	Chroococcales	*Cyanothece* sp. ATCC 51142	5,460	5,304
		Nostocales	*Nostoc punctiforme*	9,059	6,690
		Oscillatoriales	*Trichodesmium erythraeum*	7,750	4,451
		Prochlorales	*Prochlorococcus marinus*	1,670	1,921
		not defined	*Acaryochloris marina*	8,362	8,383
Proteobacteria	Alpha-proteobacteria	Caulobacterales	*Caulobacter vibrioides*	4,017	3,737
		Rhizobiales	*Rhizobium etli*	6,530	5,963
		Rhodobacterales	*Dinoroseobacter shibae*	4,418	4,187
		Rhodospirillales	*Acidiphilium cryptum*	3,963	3,559
		Rickettsiales	*Rickettsia conorii*	1,269	1,374
		Sphingomonadales	*Sphingopyxis alaskensis*	3,374	3,195
	Beta-proteobacteria	Burkholderiales	*Burkholderia mallei*	5,232	5,189
		Hydrogenophilales	*Thiobacillus denitrificans*	2,910	2,827
		Methylophilales	*Methylobacillus flagellatus*	2,972	2,753
		Neisseriales	*Neisseria meningitidis*	2,272	2,063
		Nitrosomonadales	*Nitrosomonas europaea*	2,812	2,461
		Rhodocyclales	*Aromatoleum aromaticum*	4,727	4,590
	Delta-proteobacteria	Bdellovibrionales	*Bdellovibrio bacteriovorus*	3,783	3,587
		Desulfobacterales	*Desulfotalea psychrophila*	3,660	3,234
		Desulfovibrionales	*Desulfovibrio vulgaris*	3,661	3,091
		Desulfuromonadales	*Geobacter sulfurreducens*	3,814	3,445
		Myxococcales	*Myxococcus xanthus*	9,140	7,331
		Syntrophobacterales	*Syntrophobacter fumaroxidans*	4,990	4,064
	Epsilon-proteobacteria	Campylobacterales	*Helicobacter pylori*	1,663	1,504
		Nautiliales	*Nautilia profundicola*	1,676	1,730
		not defined	*Nitratiruptor* sp. SB155-2	1,878	1,843
			Sulfurovum sp. NBC37-1	2,562	2,438
	Gamma-proteobacteria	Acidithiobacillales	*Acidithiobacillus ferrooxidans*	2,982	3,147
		Aeromonadales	*Aeromonas hydrophila*	4,744	4,122
		Alteromonadales	*Alteromonas macleodii*	4,412	4,072
		Cardiobacteriales	*Dichelobacter nodosus*	1,389	1,280
		Chromatiales	*Alkalilimnicola ehrlichei*	3,276	2,865
		Enterobacteriales	*Escherichia coli*	4,640	4,149
			Salmonella enterica	5,134	4,758
			Yersinia pestis	4,702	4,202
		Legionellales	*Legionella pneumophila*	3,576	3,206
		Methylococcales	*Methylococcus capsulatus*	3,305	2,956
		Oceanospirillales	*Chromohalobacter salexigens*	3,697	3,298
		Pasteurellales	*Pasteurella multocida*	2,257	2,015
		Pseudomonadales	*Pseudomonas aeruginosa*	6,264	5,566
		Thiotrichales	*Thiomicrospira crunogena*	2,428	2,196
		Vibrionales	*Vibrio cholerae*	4,033	3,835
		Xanthomonadales	*Xanthomonas campestris*	5,079	4,467
	not defined	not defined	*Magnetococcus* sp. MC-1	4,720	3,716
Aquificae	Aquificae	Aquificales	*Aquifex aeolicus*	1,591	1,560
Chlamydiae	Chlamydiae	Chlamydiales	*Chlamydia muridarum*	1,080	911
Verrucomicrobia	Opitutae	not defined	*Opitutus terrae*	5,958	4,612
	Verrucomicrobiae	Verrucomicrobiales	*Akkermansia muciniphila*	2,664	2,138
	not defined	not defined	*Methylacidiphilum infernorum*	2,287	2,472
Planctomycetes	Planctomycetacia	Planctomycetales	*Rhodopirellula baltica*	7,146	7,325
Spirochaetes	Spirochaetes	Spirochaetales	*Treponema pallidum*	1,138	1,036

Table. III.1 List of the bacterial genomes used in our analyses (continued)

Phylum	Class	Order	Species	Genome Size (kb)	# of Genes
Bacteroidetes	Bacteroidia	Bacteroidales	*Bacteroides fragilis*	5,241	4,231
	Flavobacteria	Flavobacteriales	*Flavobacterium johnsoniae*	6,097	5,017
	Sphingobacteria	Sphingobacteriales	*Cytophaga hutchinsonii*	4,433	3,785
Chlorobi	Chlorobia	Chlorobiales	*Chlorobaculum tepidum*	2,155	2,245
Fusobacteria	Fusobacteria	Fusobacteriales	*Fusobacterium nucleatum*	2,175	2,067
Thermotogae	Thermotogae	Thermotogales	*Thermotoga maritima*	1,861	1,858
Acidobacteria	Acidobacteria	Acidobacteriales	*Acidobacteria bacterium* Ellin345	5,650	4,777
	Solibacteres	Solibacterales	*Solibacter usitatus*	9,966	7,826
Gemmatimonadetes	Gemmatimonadetes	Gemmatimonadales	*Gemmatimonas aurantiaca*	4,637	3,935
Nitrospirae	Nitrospira	Nitrospirales	*Thermodesulfovibrio yellowstonii*	2,004	2,033
Dictyoglomi	Dictyoglomia	Dictyoglomales	*Dictyoglomus thermophilum*	1,960	1,912
Elusimicrobia	Elusimicrobia	Elusimicrobiales	*Elusimicrobium minutum*	1,644	1,529
Actinobacteria	Actinobacteria	Actinomycetales	*Mycobacterium tuberculosis*	4,412	3,989
		Bifidobacteriales	*Bifidobacterium longum*	2,260	1,729
		Rubrobacterales	*Rubrobacter xylanophilus*	3,226	3,140
Firmicutes	Bacilli	Bacillales	*Bacillus subtilis*	4,215	4,105
			Staphylococcus aureus	2,814	2,615
		Lactobacillales	*Streptococcus pneumoniae*	2,046	1,914
	Clostridia	Clostridiales	*Clostridium acetobutylicum*	4,133	3,848
		Halanaerobiales	*Halothermothrix orenii*	2,578	2,342
		Natranaerobiales	*Natranaerobius thermophilus*	3,191	2,906
		Thermoanaerobacterales	*Thermoanaerobacter*		
tengcongensis	2,689	2,588			
Tenericutes	Mollicutes	Acholeplasmatales	*Acholeplasma laidlawii*	1,497	1,380
		Entomoplasmatales	*Mesoplasma florum*	793	682
		Mycoplasmatales	*Mycoplasma pneumoniae*	816	689

3.1.2 Archaeal genomes

Of 58 currently available archaeal genomes, 18 archaeal genomes were collected so as to cover all available archaeal orders (15 archaeal orders). These genomes cover four archaeal phyla, including two major archaeal phyla, Crenarchaeota and Euryarchaeota, as well as two minor archaeal phyla, Korarchaeota (Barns *et al.*, 1996) and Nanoarchaeota (Huber *et al.*, 2002). A list of archaeal genomes used in our analyses and its taxonomy are shown in Table III.2.

3.1.3 Fungal genomes

All currently available fungal genomes were collected (15 fungal genomes). These genomes cover three fungal phyla, Ascomycota, Basidiomycota and Microsporidia, and eight fungal orders. A list of fungal genomes used in our analyses and its taxonomy are shown in Table III.3.

In order to survey how generally the correlation between protein sequence homology and gene order conservation can be observed, we selected 101 prokaryotic species for our analyses which cover all currently available prokaryotic orders. This is because

Table. III.2 List of the archaeal genomes used in our analyses

Phylum	Class	Order	Species	Genome Size (kb)	# of Genes
Crenarchaeota	Thermoprotei	Desulfurococcales	*Aeropyrum pernix*	1,670	1,700
		Nitrosopumilales	*Nitrosopumilus maritimus*	1,645	1,795
		Sulfolobales	*Sulfolobus solfataricus*	2,992	2,977
		Thermoproteales	*Pyrobaculum aerophilum*	2,222	2,605
Euryarchaeota	Archaeoglobi	Archaeoglobales	*Archaeoglobus fulgidus*	2,178	2,420
	Halobacteria	Halobacteriales	*Haloarcula marismortui*	4,275	4,240
			Halobacterium salinarum	2,571	2,622
	Methanobacteria	Methanobacteriales	*Methanothermobacter thermautotrophicus*	1,751	1,873
	Methanococci	Methanococcales	*Methanocaldococcus jannaschii*	1,740	1,786
	Methanomicrobia	Methanomicrobiales	*Methanospirillum hungatei*	3,545	3,139
		Methanosarcinales	*Methanosarcina acetivorans*	5,751	4,540
	Methanopyri	Methanopyrales	*Methanopyrus kandleri*	1,695	1,687
	Thermococci	Thermococcales	*Pyrococcus abyssi*	1,769	1,782
			Thermococcus kodakarensis	2,089	2,306
	Thermoplasmata	Thermoplasmatales	*Picrophilus torridus*	1,546	1,535
			Thermoplasma acidophilum	1,565	1,482
Korarchaeota	not defined	not defined	*Candidatus Korarchaeum cryptofilum*	1,591	1,602
Nanoarchaeota	not defined	not defined	*Nanoarchaeum equitans*	491	536

Table. III.3 List of the fungal genomes used in our analyses

Phylum	Class	Order	Species	Genome Size (kb)	# of Genes
Ascomycota	Eurotiomycetes	Eurotiales	*Aspergillus fumigatus*	29,385	9,630
			Emericella nidulans	29,699	9,410
	Sordariomycetes	Hypocreales	*Gibberella zeae*	36,354	11,628
		Sordariales	*Neurospora crassa*	37,101	10,082
	Saccharomycetes	Saccharomycetales	*Yarrowia lipolytica*	20,551	6,472
			Debaryomyces hansenii	12,250	6,334
			Eremothecium gossypii	8,766	4,722
			Kluyveromyces lactis	10,729	5,336
			Candida glabrata	12,300	5,192
			Pichia stipitis	15,441	5,816
			Saccharomyces cerevisiae	12,157	5,880
	Schizo-saccharomycetes	Schizo-saccharomycetales	*Schizosaccharomyces pombe*	12,591	5,003
Basidiomycota	Tremellomycetes	Tremellales	*Filobasidiella neoformans*	19,052	6,475
	Ustilaginomycetes	Ustilaginales	*Ustilago maydis*	19,695	6,548
Microsporidia	not defined	not defined	*Encephalitozoon cuniculi*	2,498	1,996

the computation of reciprocal all-against-all BLAST searches for all pairs of the 870 currently available prokaryotic genomes is nearly infeasible even with a high performance computing cluster system. Table III.4 shows that our collection of complete genome sequences covers a wider taxonomic space of prokaryotic species compared with the work of Lemoine *et al.* (2007).

More importantly, in order to investigate whether the finding in Dandekar *et al.* (1998) can be extended to eukaryotes, we included fungal genomes in our analyses. Although it is more desirable to include other eukaryotes such as animals and plants as well as fungi, it requires a more complicated (or sophisticated) workflow to detect conserved gene clusters because higher eukaryotic genomes have gone through nu-

Table. III.4 Taxonomic space covered by our analyses and the work of Lemoine *et al.* (2007)

Domain	Rank	Lemoine *et al.* (2007)	Our analyses
Bacteria	Phylum	14	21
	Class	20	35
	Order	43	79
Archea	Phylum	3	4
	Class	10	11
	Order	12	15
Fungi	Phylum	0	3
	Class	0	7
	Order	0	8

merous tandem duplication events. Thus we analyzed only fungal genomes regarding eukaryotic genomes in the present study, although we have a plan to improve the OASYS algorithm so as to be able to accurately identify OGs even when there exist tandem duplications and to examine whether the correlation between protein sequence homology and gene order conservation can be observed also in higher eukaryotes.

3.2 Workflow for detecting conserved gene clusters

A conserved gene cluster is defined as a cluster of neighboring genes whose gene order is conserved across several species. Detecting conserved gene clusters between pairwise genomes is one of the most important steps in our analyses. Our purpose is to compare evolutionary distances separating orthologous genes (OGs) from two organisms between OGs in conserved gene clusters (clustered OGs) and OGs that are not the members of conserved gene clusters (isolated OGs). Thus, both accurate identification of orthology relationships and accurate detection of conserved gene clusters are necessary to ensure that the differences between clustered OGs and isolated OGs are not the artifacts caused by inaccurate workflow.

A difficulty in the identification of OGs is associated with the discrimination between orthologs, which are genes evolved by vertical descent from a single ancestral gene, and paralogs, which are genes evolved by duplication (Fitch, 1970). Given a timing of the speciation separating two genomes, paralogs that go through duplication events after the speciation are referred to as in-paralogs, whereas paralogs that are duplicated before the speciation are referred to as out-paralogs (Remm *et al.*, 2001). In many cases where in-paralogs exist, similarity of protein sequences is not sufficient information to determine which of the in-paralogs is functionally equivalent to the ortholog in the other species. Due to this uncertainty of functional equivalence between in-paralogs, the vast majority of recently proposed methods identify many-to-many orthology relationships, i.e. all of in-paralogs are clustered together in an orthologous group (Remm *et al.*, 2001; Li *et al.*, 2003; Tatusov *et al.*, 2003; Dehal and Boore, 2006;

Vilella *et al.*, 2009).

However, in-paralogs could be under different evolutionary pressures. Evolutionary biologists consider that one of the in-paralogs have retained the ancestral function and the other in-paralogs have acquired new lineage-specific functions. Thus, the one of in-paralogs would be under the evolutionary pressures to maintain protein sequences, whereas the others would not be (Ohno, 1970; Zhang *et al.*, 1998; Moore and Purugganan, 2003; Rodriguez-Trelles *et al.*, 2003; Thornton and Long, 2005; Han *et al.*, 2009). In order to focus on the correlation between protein sequence homology and gene order conservation, and to exclude the undesirable effects of in-paralogs, our workflow identifies one-to-one orthology relationships rather than many-to-many. Even when there exist in-paralogs, our workflow identifies one-to-one orthology relationships by selecting the orthologous gene pairs that are located on the corresponding chromosomal positions. Since such OGs tend to have retained the ancestral function (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999a,b; Snel *et al.*, 2000; Notebaart *et al.*, 2005), the OGs identified by our workflow would be less affected by in-paralogs.

Our workflow starts with parsing GenBank files. Subsequently, one-to-one orthology relationships of genes are identified by the OASYS program. Thereafter, OGs that are strictly adjacent in both genomes are clustered together in order to detect conserved gene clusters, in which neither insertion/deletion of genes nor inversion is allowed. This clustering criterion is the same as the work of Lemoine *et al.* (2007).

An originality of our workflow is to use the information of gene order conservation in the step to identify OGs. Suppose that two genomes, G_A and G_B, have evolved from a common ancestor, and the gene order of three neighboring genes have not been disrupted. Let the descendant of the gene cluster in G_A and G_B be $\{a_{i-1}, a_i, a_{i+1}\}$ and $\{b_{i-1}, b_i, b_{i+1}\}$, respectively. In addition, suppose that b_i is duplicated after the speciation of G_A and G_B, and G_B comes to encode a new gene b'_i as shown in Fig. III.9. In this case, a heuristic homology search tool like BLAST might yield a smaller similarity score for the gene pair (a_i, b_i) than the gene pair (a_i, b'_i) even though the gene pair (a_i, b_i) be truly orthologous. Then, the gene pair (a_i, b_i) would not be identified as orthologous by the methods based only on protein sequences and therefore the conserved gene cluster of $\{a_{i-1}, a_i, a_{i+1}\}$ and $\{b_{i-1}, b_i, b_{i+1}\}$ would not be detected. On the other hand, the information of gene order conservation enhances to identify three one-to-one orthology relationships, (a_{i-1}, b_{i-1}), (a_i, b_i), and (a_{i+1}, b_{i+1}), which would yield the detection of the conserved gene cluster of the three OGs. Accordingly, in order to sensitively detect conserved gene clusters even if there exist in-paralogs, OGs are need to be identified based not only on the information of protein sequences but also on the information of gene order conservation.

3.2.1 Parsing GenBank files

We used the Bio::SeqIO module in the BioPerl package (Stajich *et al.*, 2002) to parse GenBank files drawn from the NCBI FTP server. For each CDS feature in a GenBank file, we extracted the locus tag, protein sequence, chromosomal positions of coding sequences, and genetic code that is used to translate the coding sequences. The coding

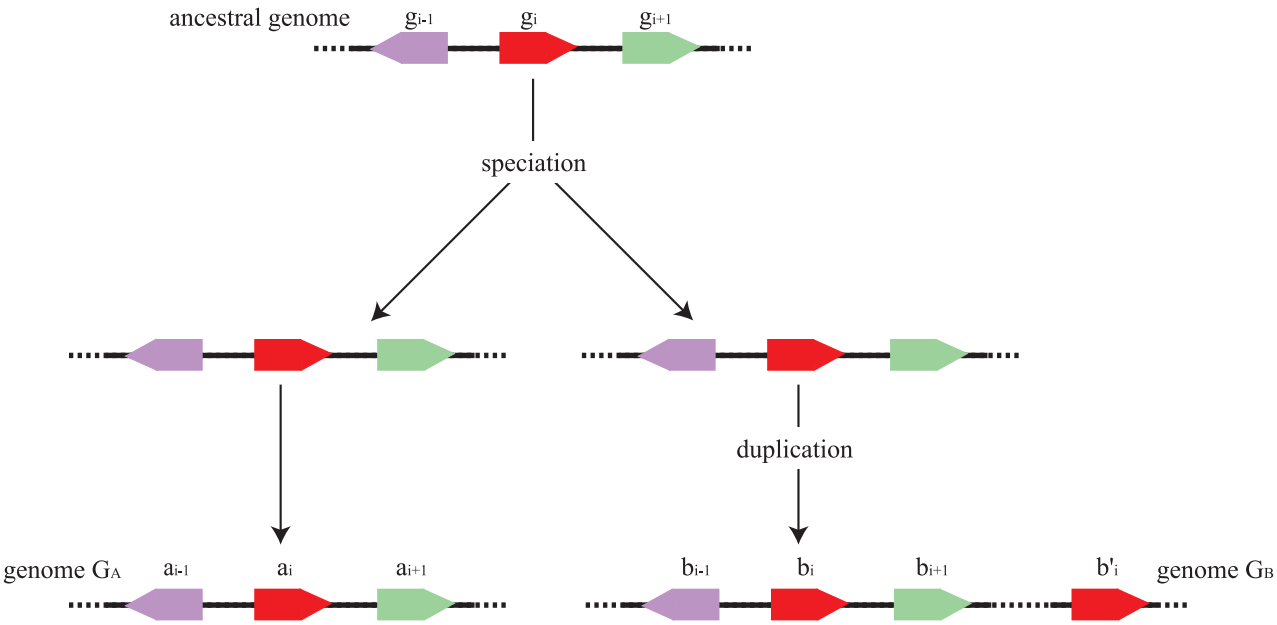


Fig. III.9 An illustration of a genome evolution with a duplication event. We here suppose that two genomes, G_A and G_B, are speciated from a common ancestor, and the gene order of three neighboring genes have not been disrupted. The descendant of the gene cluster in G_A and G_B are denoted as $\{a_{i-1}, a_i, a_{i+1}\}$ and $\{b_{i-1}, b_i, b_{i+1}\}$, respectively. In addition, we suppose that b_i is duplicated after the speciation of G_A and G_B, and G_B comes to encode a new gene b'_i.

sequence for the CDS feature was obtained by extracting the DNA sequences from the whole genome sequence described in the GenBank file by using the chromosomal positions of coding sequences. Then, we assigned our unique gene ID to the CDS, and the DNA sequence, protein sequence, chromosomal position, and genetic code were associated with the gene ID.

3.2.2 Identifying orthologous genes

A file containing all protein sequences was created for each organism. Subsequently, we executed reciprocal all-against-all BLAST searches by using the BLASTP program (Altschul *et al.*, 1990) with default parameters. Suspicious BLAST hits were filtered out by eliminating the BLAST hits whose bit score is lower than 50 bits and the BLAST hits whose matching segment is shorter than the half length of the protein sequences. Then, we used the OASYS program (version 0.2) with default parameters to identify one-to-one orthology relationships of genes.

3.2.3 Detecting conserved gene clusters

In order to detect conserved gene clusters, we input the results of the OASYS program into the dpd clustering program included in the OASYS distribution. In this computation, the threshold of the distance between OGs to cluster together was set at 1.0. By doing so, only the strictly adjacent OGs are clustered together.

3.2.4 Computing PAM distance

Given two protein sequences, we computed the global alignment of the two sequences by using the needle program included in the EMBOSS package (version 6.0.1) (Rice *et al.*, 2000). This computation was executed with default parameters. Subsequently, the PAM distance separating the two protein sequences was computed by using protdist program included in the Phylip package (version 3.68) (Felsenstein, 2005). This computation was executed with default parameters except for setting the model at the Dayhoff PAM matrix. In this setting, the DCMut model (Kosiol and Goldman, 2005) was used to compute PAM distances.

3.2.5 Estimating K_A and K_S values

In order to obtain the alignment of two coding sequences, we reused the alignment of two protein sequences, which had been calculated to compute PAM distances. The alignment of two DNA sequences were simply calculated by matching protein sequences and DNA sequences. Thereafter, we used the yn00 program included in the PAML package (version 4.2) (Yang, 1997) to estimate the K_A and K_S values. The yn00 program is an implementation of the algorithm proposed by Yang and Nielsen (2000), which takes into account transition/transversion rate bias and base/codon frequency bias. In this computation, the yn00 program was executed with default parameters except for setting the icode parameter at the genetic code of the input coding sequences. Since several genetic codes cannot be analyzed by the yn00 program, we modified the program so that all genetic codes accepted by NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi) can be analyzed. To our knowledge, there is no appropriate method to compute K_A and K_S values in the case where the genetic codes of two coding sequences are different. Accordingly, we could neither compute K_A and K_S values nor conduct further analyses in such cases.

4 Results and Discussion

4.1 Validation of our workflow

In order to validate the effectiveness of our workflow, we compared the results of our workflow with three alternative approaches. The first alternative approach is the method used in Lemoine *et al.* (2007). They applied the RSD (reciprocal smallest distance) method (Wall *et al.*, 2003) to the identification of putative OGs, and additional orthologs (in-paralogs) are detected by reconstructing a phylogenetic tree for each gene family and by using an *ad hoc* algorithm to determine whether each internal node of the phylogenetic tree corresponds to a speciation event, or a duplication event. The union of the OGs obtained from the RSD method and the OGs obtained from the *ad hoc* phylogenetic approach is used to detect conserved gene clusters. Accordingly, the Lemoine's method allows many-to-many orthology relationships and might detect false positives. To avoid detecting false positives, they used very strict cutoff criteria

Table. III.5 Number of orthologous gene pairs identified by four alternative approaches (Number of clustered OGs / Total number of OGs)

E.coli proteome compared with	Lemoine *et al.* (2007)	RBH	Syntenator	OASYS
S. enterica	700 / 2,592 (27%)	2,737 / 3,003 (91%)	2,378 / 2,511 (95%)	2,768 / 3,014 (92%)
B. subtilis	229 / 994 (23%)	225 / 1,090 (21%)	142 / 155 (92%)	272 / 1,100 (25%)
B. thetaiotaomicron	128 / 802 (16%)	124 / 893 (15%)	64 / 64 (100%)	152 / 898 (17%)
M. acetivorans	60 / 431 (14%)	48 / 518 (9%)	77 / 77 (100%)	65 / 537 (12%)

Table. III.6 Statistics of conserved gene clusters detected by four alternative approaches

	Lemoine *et al.* (2007)		RBH		Syntenator		OASYS	
E.coli proteome compared with	No. of Clusters[a,c]	Max size[b]	No. of Clusters[a]	Max size[b]	No. of Clusters[a]	Max size[b]	No. of Clusters[a]	Max size[b]
S. enterica	-	20	431	39	346	44	429	44
V. cholerae	-	10	318	22	107	22	330	22
P. aeruginosa	-	12	250	22	94	22	267	22
M. loti	-	9	112	9	102	8	141	9
B. subtilis	-	9	92	10	40	10	110	10
M. tuberculosis	-	6	52	9	10	5	62	9
C. tepidum	-	9	54	10	8	22	59	10
M. acetivorans	-	3	22	4	17	5	30	4
S. solfataricus	-	3	12	3	7	4	14	3

[a] Number of conserved gene clusters detected by each method.
[b] Maximum size of conserved gene clusters detected by each method.
[c] Since the detailed results of the work of Lemoine *et al.* (2007) are not available, the number of conserved gene clusters cannot be examined.

to filter out homologous gene pairs. The second and third alternative approaches use the RBH (reciprocal best hit) method (Tatusov *et al.*, 1997) and the Syntenator program (Rödelsperger and Dieterich, 2008) to identify OGs, respectively. The RBH method is a well-known method to identify OGs and is based only on similarities of protein sequences. Meanwhile, Syntenator identifies OGs by simultaneously finding conserved gene orders, and therefore is based not only on the information of protein sequence homology but also the information of gene order conservation. In our workflow and alternative approaches, the same algorithm is applied to the clustering of adjacent OGs. Note that we applied our clustering algorithm also to the OGs identified by Syntenator even though Syntenator detects not only OGs but also conserved gene orders because the definition of conserved gene orders in Syntenator allows insertions/deletions of genes, and is slightly different from our definition of conserved gene clusters. The differences in the results of conserved gene clusters among alternative approaches directly reflect the differences in the algorithm to identify OGs.

Table III.5 shows the number of the OGs identified by the four alternative methods, as well as the number of clustered OGs. Table III.6 summarizes the statistics of the conserved gene clusters detected by the four alternative approaches. We can see in Table III.6 that our workflow tends to detect a larger number of conserved gene clusters compared with the RBH approach, and the maximum size of the conserved gene clusters detected by our workflow was greater than the method used in Lemoine *et al.* (2007). Moreover, the histograms of the size of conserved gene clusters show that

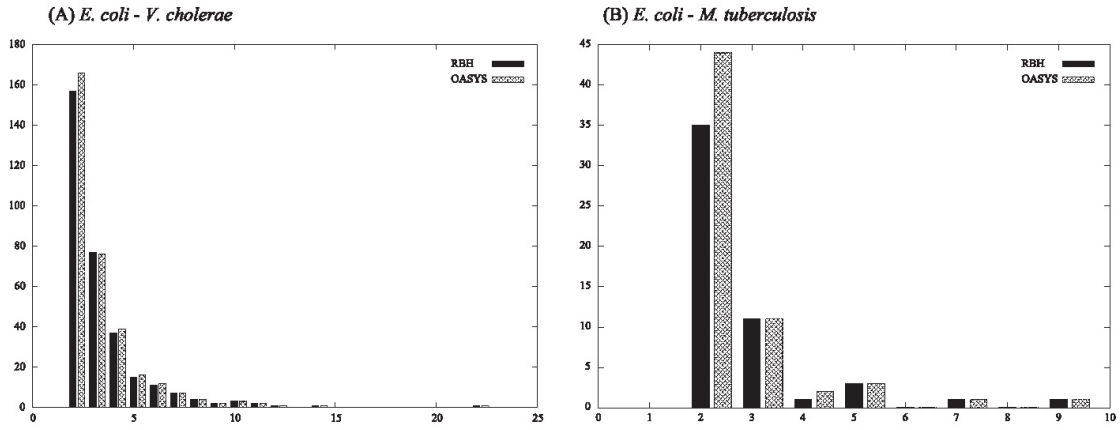


Fig. III.10 Histogram of the size of conserved gene clusters. The sizes of the conserved gene clusters detected by the RBH method and our workflow (OASYS) are compared. (A) Histogram of the size of conserved gene clusters detected by comparing *E. coli* and *V. cholerae*. (B) Histogram of the size of conserved gene clusters detected by comparing *E. coli* and *M. tuberculosis*.

the difference in the number of conserved gene clusters between the RBH method and our workflow is mostly due to the difference in the number of conserved gene clusters whose size is two (Fig. III.10), indicating that our workflow enables sensitive detection of small conserved gene clusters. Thanks to this sensitiveness, our workflow could detect a larger number of clustered OGs than the other three methods (Table III.5). Table III.5 also shows that the number of OGs detected by our workflow is a little greater than the RBH method and largely greater than the Lemoine's method. This result indicates that a number of *bona fide* OGs are missed by the Lemoine's method, possibly due to very strict cutoff criteria to filter out homologous gene pairs. Our workflow avoids false positives of OGs by using the information of gene order conservation in order to distinguish genuine orthologous gene pairs from the other homologous gene pairs, and therefore, does not need to use such too stringent cutoff criteria.

As an example of the differences between the RBH method and our workflow, we here focus on a conserved gene cluster detected between *E. coli* and *M. tuberculosis*, which is composed of sulfate and thiosulfate transport genes. As illustrated in Fig. III.11, our workflow detects the conserved gene cluster composed of four OGs, whereas the workflow based on the RBH method detects the conserved gene cluster composed of three OGs. This is because our workflow identifies *cysP* as the ortholog of *subI*, on the other hand, the RBH method identifies *sbp*. In *E. coli*, mutation experiments and presumption based on sequence homology suggest that CysP, CysU, CysW and CysA form a complex of sulfate/thiosulfate ABC transporter, and mRNAs of these subunits are cotranscribed (Hryniewicz *et al.*, 1990; Sirko *et al.*, 1990). Also in *M. tuberculosis*, *subI*, *cysT*, *cysW*, and *cysA1* are predicted to constitute an

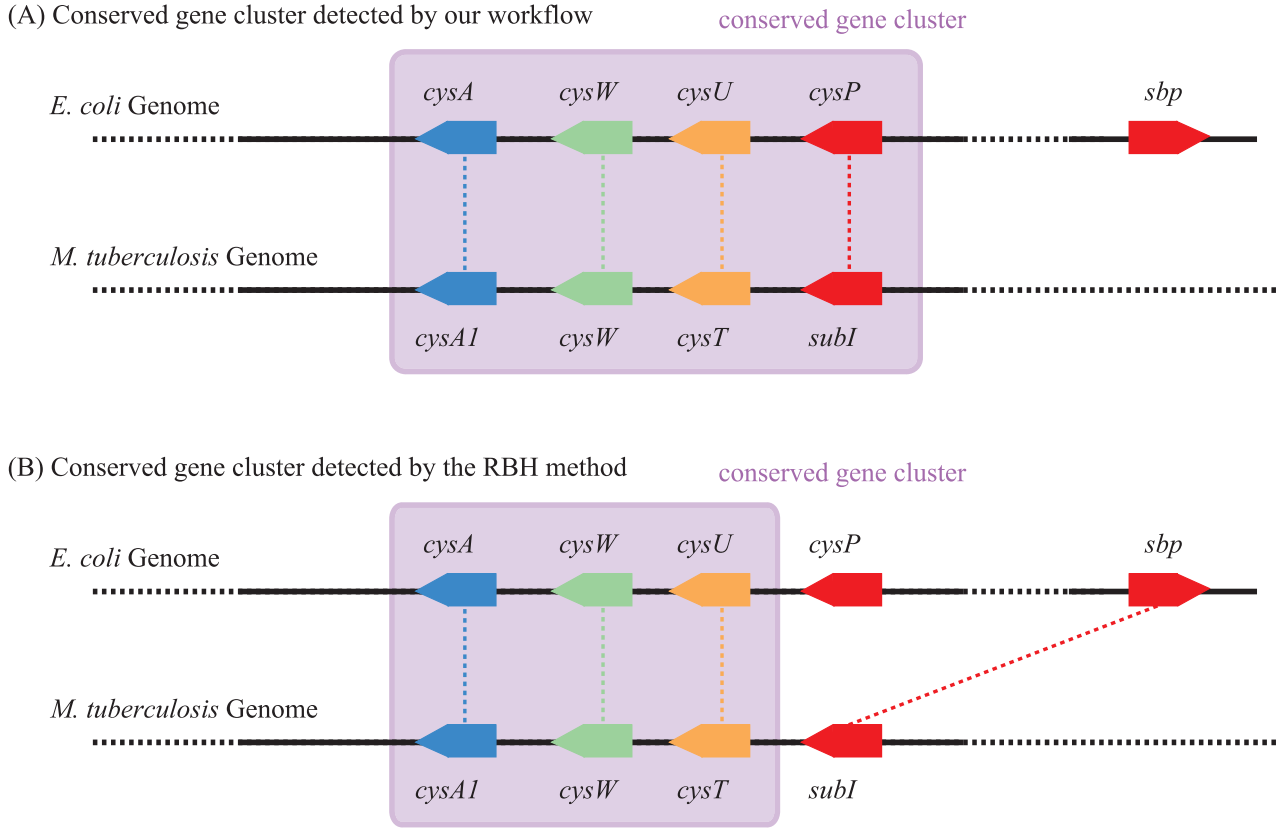


Fig. III.11 Conserved gene cluster of sulfate and thiosulfate transport genes. Colored arrows represent genes, and homologous genes are depicted as the arrows having the same color, e.g. both *cysA* in *E. coli* and *cysA1* in *M. tuberculosis* are blue-colored, representing that the two genes are homologous. Orthologous genes detected by each method are connected by colored broken lines. Conserved gene clusters are depicted as colored blocks. (A) A conserved gene cluster detected by our workflow is illustrated. Four orthologous gene pairs, (*cysA*, *cysA1*), (*cysW*, *cysW*), (*cysU*, *cysT*), and (*cysP*, *subI*), were detected by our workflow. The clustering of neighboring OGs results in a conserved gene cluster whose size is four. (B) A conserved gene cluster detected by the RBH method is illustrated. Four orthologous gene pairs, (*cysA*, *cysA1*), (*cysW*, *cysW*), (*cysU*, *cysT*), and (*sbp*, *subI*), were detected by the RBH method. The clustering of neighboring OGs yields a conserved gene cluster whose size is three.

operon (Alm *et al.*, 2005; Price *et al.*, 2005). Taken together, *subI* in *M. tuberculosis* seems to play an equivalent role as *cysP* in *E. coli*. This example indicates that our workflow can correctly identify *bone fide* OGs by taking into account the information of gene order conservation, and demonstrates that our workflow can avoid underestimating the size of conserved gene clusters.

Compared with the Syntenator program, OASYS detects a larger number of conserved gene clusters while the maximum size of conserved gene clusters tends to be smaller in the comparisons of distantly related genomes (Table III.6). The number of OGs and clustered OGs detected by OASYS were consistently greater than

Table. III.7 Results of our workflow in the comparison of prokaryotic genomes

	No. of genome pairs			
	All[a]	# of OGs $\geq$ 300[b]	%Clustered $\geq$ 10%[c]	(No. of OGs $\geq$ 300) AND (%Clustered $\geq$ 10%)[d]
bacteria-bacteria comparisons	3,403	3,204 (94.2%)	3,342 (98.2%)	3,143 (92.4%)
archaea-archaea comparisons	153	136 (88.9%)	135 (88.2%)	135 (88.2%)
bacteria-archaea comparisons	1,494	812 (54.4%)	722 (48.3%)	497 (33.3%)
Total	5,050	4,152 (82.2%)	4,199 (83.1%)	3,775 (74.8%)

[a]Number of all pairwise combinations of genomes.
[b]Number of genome pairs where more than 300 OGs were identified.
[c]Number of genome pairs where the percentage of OGs in conserved gene clusters exceeded 10%.
[d]Number of genome pairs where more than 300 OGs were identified and the percentage of OGs in conserved gene clusters exceeded 10%.

Syntenator, although the percentages of OGs in conserved gene clusters detected by Syntenator were greater than those of OASYS (Table III.5). These results indicate that the advantage of OASYS over Syntenator lies in the sensitivity to identify both clustered and isolated OGs, which can be accomplished by detecting small conserved gene clusters sensitively, whereas Syntenator is suitable to detect large conserved gene clusters especially in the comparisons of remotely related genomes. From the point of view that OGs will be used to statistically test the differences between clustered and isolated OGs in our analyses, it is needed to detect isolated OGs sensitively as well as clustered OGs, and therefore, OASYS is more appropriate for our analyses than Syntenator.

4.2 Results of comparing prokaryotic genomes

We applied our workflow to all pairwise combinations of the 101 prokaryotic (83 bacterial and 18 archaeal) genomes listed in Tables III.1 and III.2, and one-to-one orthology relationships of genes and conserved gene clusters were computed for each pair of genomes. The number of OGs and the percentage of OGs in conserved gene clusters are visualized in Figs. III.12 and III.13, respectively, and these results are summarized in Table III.7. We can see in Table III.7 that the percentage of OGs in conserved gene clusters exceeded 10% in almost cases of bacteria-bacteria genome comparisons (98.2%) and archaea-archaea genome comparisons (88.2%). Even when comparing bacterial and archaeal genomes, for 722 of 1,494 genome pairs (48.3%), the percentage of OGs in conserved gene clusters exceeded 10%, indicating that local gene orders are substantially conserved even between bacterial and archaeal genomes.

Further sequence analyses were conducted for the genome pairs where more than 300 OGs were detected and the percentage of OGs in conserved gene clusters exceeded 10%. First, the PAM distance, which is the number of accepted point mutations per 100 residues, were computed for each orthologous gene pair. Then, we examined whether the PAM distances of clustered OGs are significantly lower than those of isolated OGs. In our statistical test, the null hypothesis (H_0) assumes that the population distribution of the PAM distances of clustered OGs is identical to the population

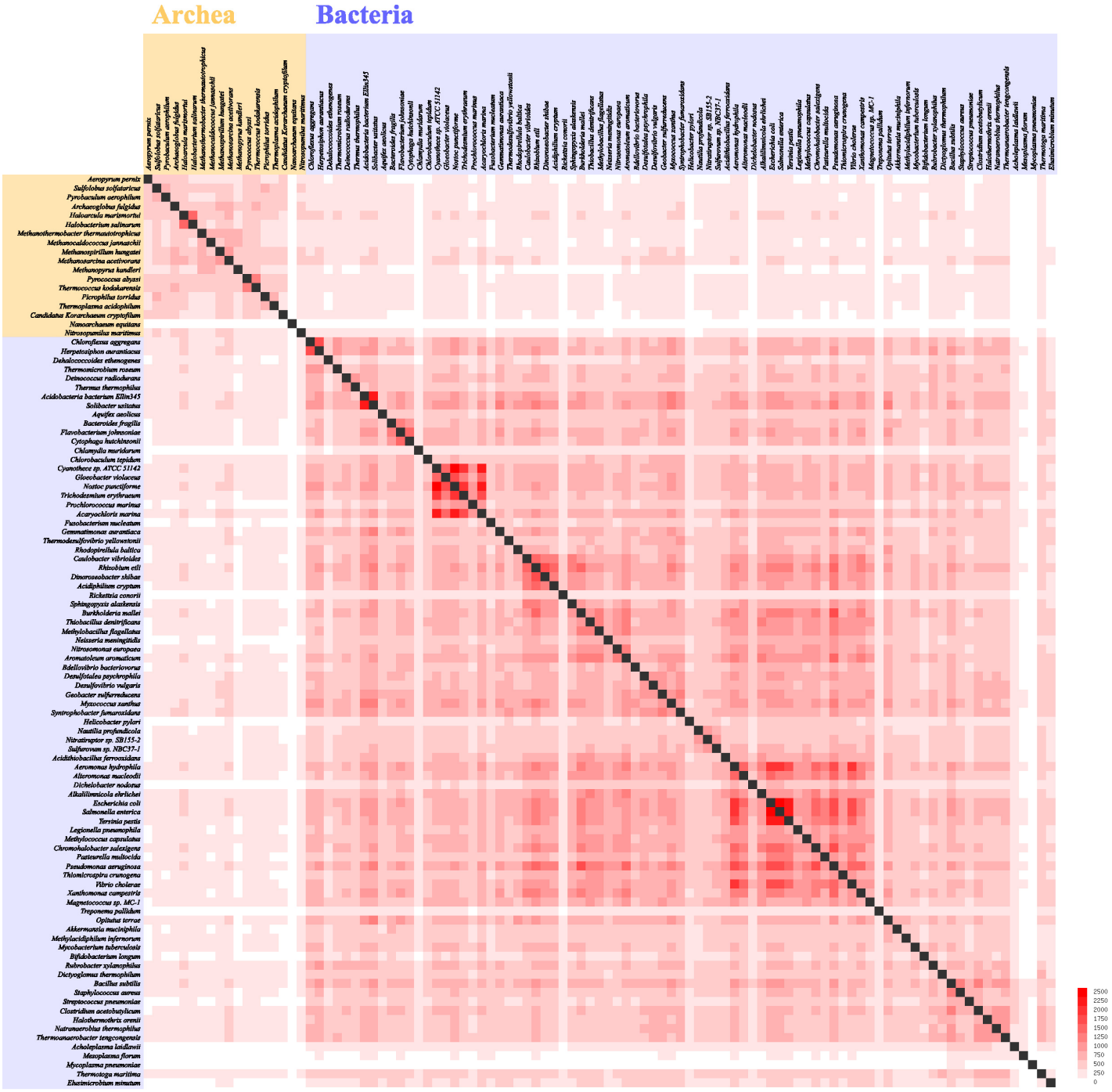


Fig. III.12 Number of OGs identified in the comparisons of prokaryotic genomes. A column or a row in this matrix corresponds to a prokaryotic organism, and the result of comparing two prokaryotic organisms is shown in the corresponding cell. The color of each cell represents the degree of the number of OGs identified by our workflow. Analyses corresponding to black-colored cells were not conducted.

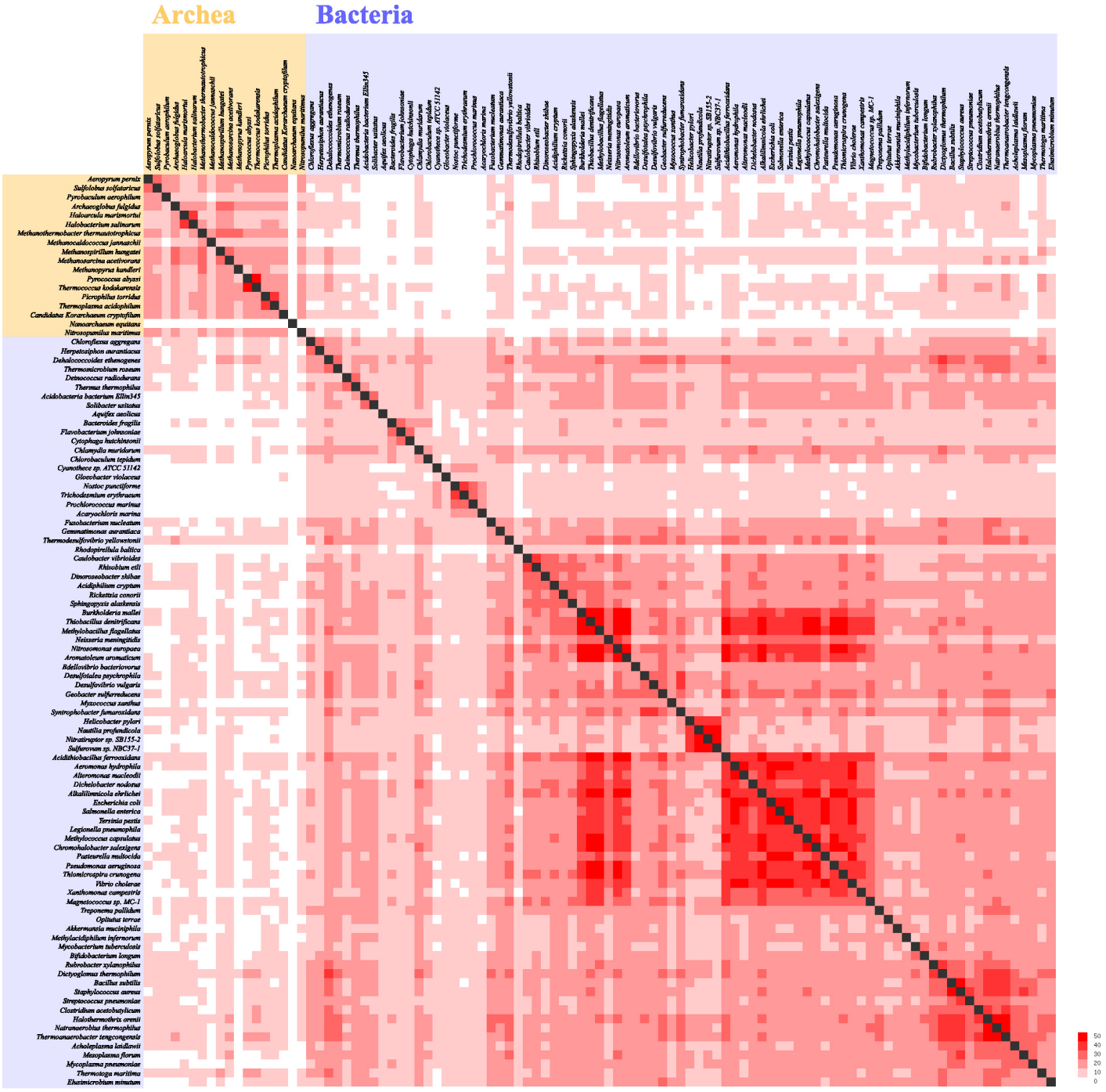


Fig. III.13 Percentage of OGs in conserved gene clusters. A column or a row in this matrix corresponds to a prokaryotic organism, and the result of comparing two prokaryotic organisms is shown in the corresponding cell. The color of each cell represents the degree of percentage of OGs in conserved gene clusters. Analyses corresponding to black-colored cells were not conducted.

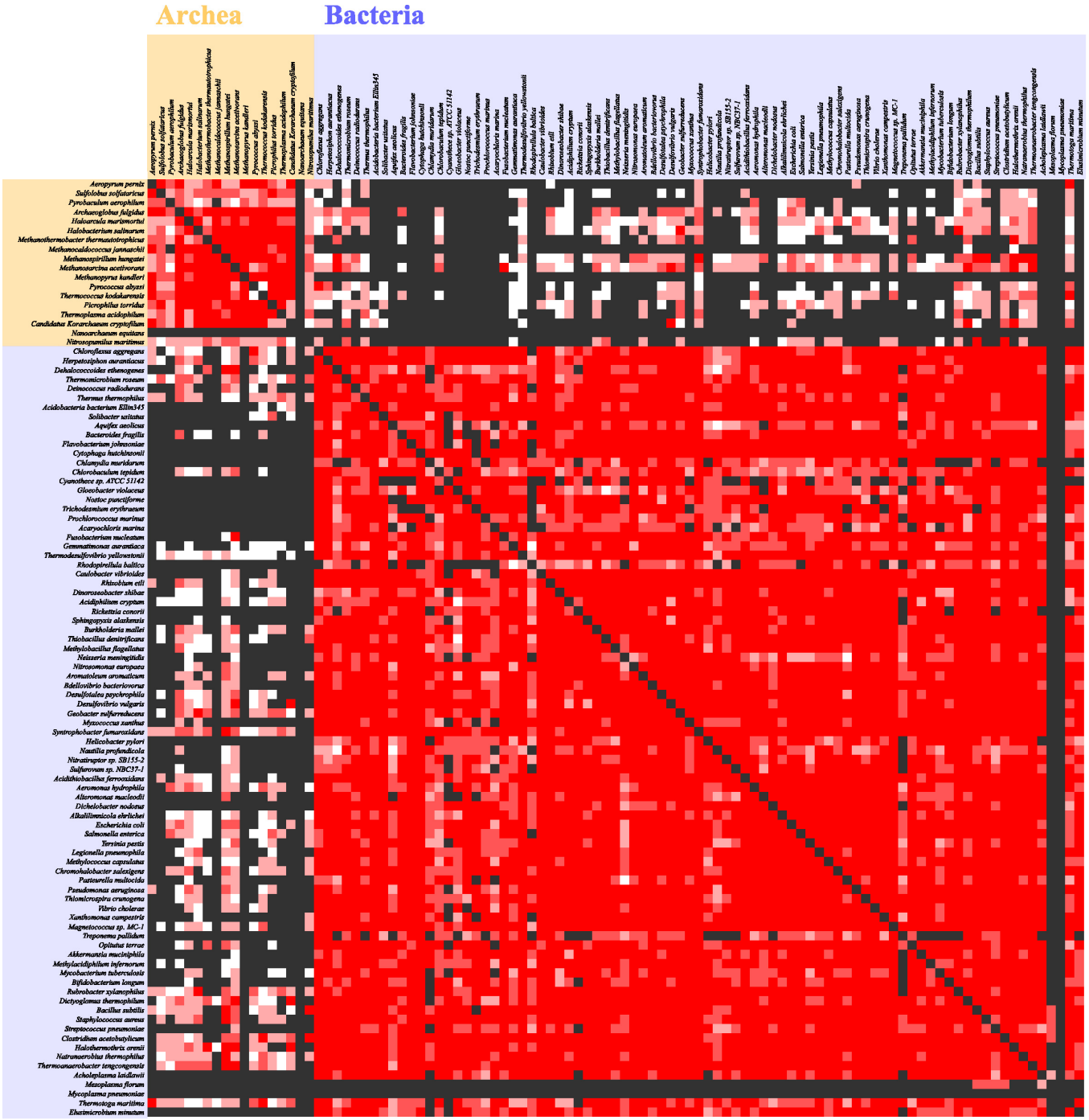


Fig. III.14 p-value of the difference in PAM distance. A column or a row in this matrix corresponds to a prokaryotic organism, and the result of comparing two prokaryotic organisms is shown in the corresponding cell. The color of each cell represents the degree of the logarithm (base 10) of the p-value of the difference in PAM distance. Analyses corresponding to black-colored cells were not conducted.

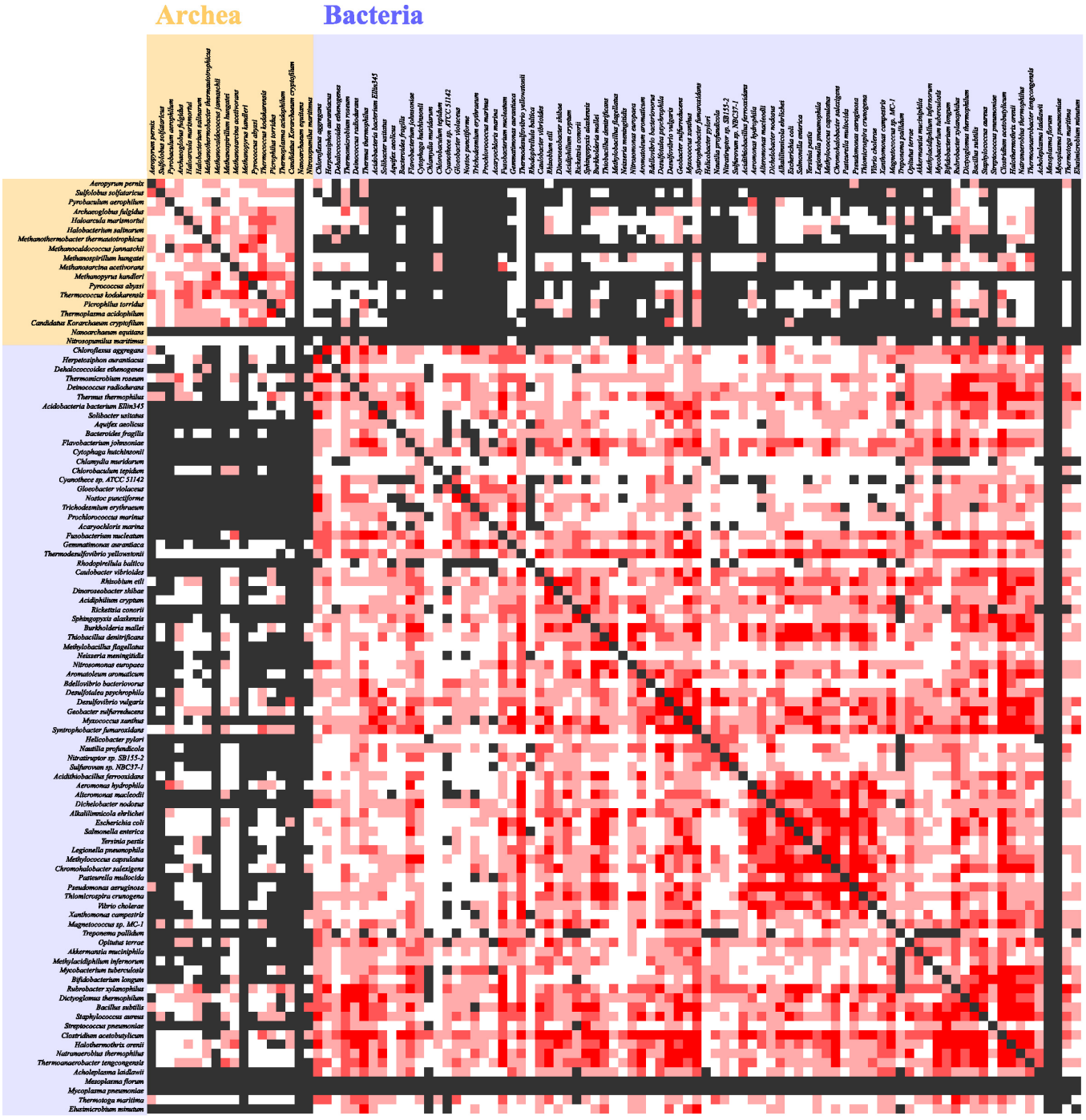


Fig. III.15 p-value of the difference in K_A/K_S ratio. A column or a row in this matrix corresponds to a prokaryotic organism, and the result of comparing two prokaryotic organisms is shown in the corresponding cell. The color of each cell represents the degree of the logarithm (base 10) of the p-value of the difference in K_A/K_S ratio. Analyses corresponding to black-colored cells were not conducted.

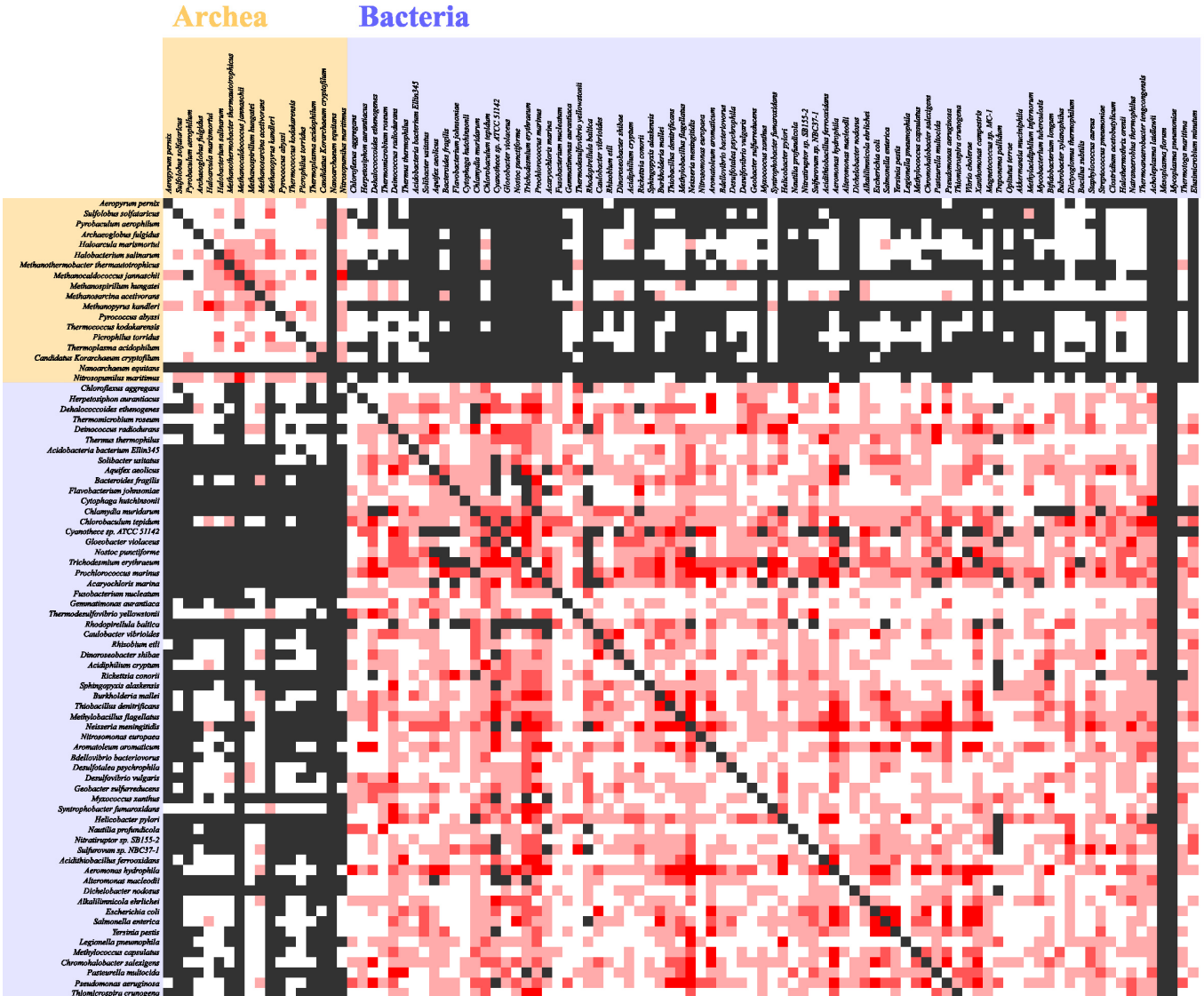


Fig. III.16 p-value of the difference in K_S value. A column or a row in this matrix corresponds to a prokaryotic organism, and the result of comparing two prokaryotic organisms is shown in the corresponding cell. The color of each cell represents the degree of the logarithm (base 10) of the p-value of the difference in K_S value. Analyses corresponding to black-colored cells were not conducted.

Table. III.8 Number of genome pairs that show significant difference

	# of genome pairs				
	PAM distance[a]	K_A/K_S and K_S[b]	K_A/K_S[c]	K_S[d]	no difference[e]
bacteria-bacteria comparisons	3,137	856 (27.3%)	1,157 (36.9%)	883 (28.1%)	241 (7.7%)
archaea-archaea comparisons	130	19 (14.6%)	51 (39.2%)	28 (21.5%)	32 (24.6%)
bacteria-archaea comparisons	322	0 (0.0%)	109 (33.9%)	18 (5.6%)	195 (60.1%)
Total	3,589	875 (24.4%)	1,317 (37.0%)	929 (25.9%)	468 (13.0%)

[a]Number of genome pairs that show a significant difference in PAM distance between clustered OGs and isolated OGs.
[b]Number of genome pairs that show a significant difference both in K_A/K_S ratio and in K_S value.
[c]Number of genome pairs where a significant difference in K_A/K_S ratio was detected but no significant difference in K_S value was observed.
[d]Number of genome pairs where a significant difference in K_S value was detected but no significant difference in K_A/K_S ratio was observed.
[e]Number of genome pairs that do not show any significant difference neither in K_A/K_S ratio nor K_S value.

distribution of the PAM distances of isolated OGs. The alternative hypothesis (H_1) assumes that the population distribution of the PAM distances of clustered OGs has a smaller mean than that of isolated OGs. Since the population of PAM distances cannot be assumed to be normally distributed, Mann-Whitney U-test (Wilcoxon, 1945; Mann and Whitney, 1947) was employed to compute the p-values. Fig. III.14 visualizes the p-value computed for each pair of genomes, and the results are summarized in Table III.8. Of 3,143 bacterial genome pairs analyzed, significant difference in PAM distance was detected for 3,137 genome pairs (99.8%) with the p-value cutoff at 0.01. Of 135 archaeal genome pairs analyzed, significant difference in PAM distance was detected for 130 genome pairs (96.3%). These results confirm the previous finding in Dandekar *et al.* (1998) that the degree of protein sequence conservation of clustered OGs is substantially higher than that of isolated OGs. Moreover, the significant difference in PAM distance was observed for 322 genome pairs of bacterial and archaeal genomes (64.8%), suggesting that the finding of Dandekar *et al.* (1998) is a general trend among prokaryotic genomes.

In order to shed light on the evolutionary forces behind the correlation between protein sequence homology and gene order conservation, we estimated the rate of synonymous substitutions (K_S) and the rate of nonsynonymous substitutions (K_A) for each orthologous gene pair. Subsequently, we conducted statistical tests to assess whether the K_A/K_S ratio (or K_S value) of clustered OGs is significantly lower than that of isolated OGs. In these statistical tests, the null and alternative hypotheses are assumed in a similar manner to the statistical tests for the difference in PAM distance. Figs. III.15 and III.16 visualize the p-values computed for each pair of genomes, and the results are summarized in Table III.8. We can see in Table III.8 that, of 3,589 prokaryotic genome pairs that show the significant difference in PAM distance, significant difference was detected both in K_A/K_S ratio and in K_S value

for 875 genome pairs (24.4%). For 1,317 prokaryotic genome pairs (37.0%), there were significant differences in K_A/K_S ratio, but no significant difference in K_S value. For 883 prokaryotic genome pairs (28.1%), significant differences in K_S value were observed, but no significant difference in K_A/K_S ratio was detected. These results interestingly indicate that although the correlation between protein sequence homology and gene order conservation is consistently observed and seems to be a general trend among prokaryotic genomes, the underlying mechanisms behind the correlation are different among lineages.

Dandekar *et al.* (1998) postulates a hypothesis for the underlying mechanism and explains why the gene order conservation can be useful to predict gene functions from the point of view of co-adaptation (Fisher, 1930; Wallace, 1991; Pazos and Valencia, 2008). Proteins that interact physically tend to be co-adapted, and co-adapted genes would be under positive selection to form clusters of co-adapted genes and/or selective pressures to maintain gene clusters in order to reduce the chance of genetic recombination perturbing co-adapted pairs of genes. Moreover, genes whose products interact physically should exhibit a lower rate of mutation, because of the selective constraints imposed by the interaction. Taken together, gene order conservation should correlate with protein sequence homology and the interaction of proteins. Our results provide an impact on the hypothesis because there are cases that higher degree of protein sequence conservation would be caused by lower substitution rate of coding sequences rather than stronger selective pressures to preserve protein sequences, which cannot be explained by the hypothesis of Dandekar *et al.* (1998). Thus, our finding requires another hypothesis for the underlying mechanisms that yield the correlation between protein sequence homology and gene order conservation. For example, we can explain the correlation from the point of view of regional variation in mutation rates (Wolfe *et al.*, 1989; Baer *et al.*, 2007). Though neutral mutation rates were once considered to be uniform along with chromosomes, it has been discovered in multicellular organisms that they can vary among segmental regions of a single chromosome (Baer *et al.*, 2007; Fox *et al.*, 2008). Moreover, it has been reported that the rate of nucleotide substitutions for each segmental region is correlated with the recombination rate in eutherian genomes (Hardison *et al.*, 2003). We postulate that the rate of nucleotide substitutions might be correlated with the recombination rate and/or the rearrangement rate (Sémon and Wolfe, 2007) also in prokaryotic genomes, and such correlation could yield the correlation between protein sequence homology and gene order conservation.

4.3 Results of comparing fungal genomes

We applied our workflow to all pairwise combinations of the 15 fungal genomes listed in Table III.3, and one-to-one orthology relationships of genes and conserved gene clusters were computed for each pair of genomes. The number of OGs and the percentage of OGs in conserved gene clusters are visualized in Figs. III.17A and III.17B, respectively. These figures show that more than 1,000 OGs were detected even between distantly related fungal genomes (Fig. III.17A), whereas the percentage of OGs in

conserved gene clusters did not exceed 10% when comparing fungal genomes across classes (Fig. III.17B), suggesting that extensive gene shuffling has been occurred during fungal genome evolution.

We conducted further sequence analyses for the genome pairs, where more than 500 OGs were identified and the percentage of OGs in conserved gene clusters exceeded 10%. Similar to the analyses of prokaryotic genomes, the difference in PAM distance between clustered and isolated OGs was statistically tested for each pair of fungal genomes, and the results are visualized in Fig. III.17C. To our surprise, the significant differences were observed in more than half of fungal genome pairs. Especially in the comparison of genomes in the subphylum Pezizomycotina, strongly significant difference was observed. In order to demonstrate that the correlation between protein sequence homology and gene order conservation observed in fungal genomes is independent of the algorithm of OASYS, we examined whether the correlation can be detected by an alternative approach. We identified OGs between *A. fumigatus* and *A. nidulans* by using the Syntenator program (Rödelsperger and Dieterich, 2008), and the OGs identified were clustered by the dpd clustering program in the OASYS distribution. A Mann-Whitney U-test showed a significant difference in PAM distance between clustered and isolated OGs (p-value $\leq 1.28 \times 10^{-20}$). We also computed the p-value in the comparison of *G. zeae* and *N. crassa*, and a significant difference was observed (p-value $\leq 6.77 \times 10^{-3}$). These results indicate that the correlation between protein sequence homology and gene order conservation observed in fungal genomes is independent of our workflow and would be a genuine trend in fungal genomes.

In order to survey the evolutionary forces behind the correlation observed in fungal genomes, the differences in K_A/K_S ratio and K_S value between clustered and isolated OGs were statistically tested (Figs. III.17D and III.17E). Fig. III.17D shows that strong significant difference in K_A/K_S ratio was observed when comparing genomes in the subphylum Pezizomycotina, whereas no significant difference in K_A/K_S ratio was detected when comparing genomes in the class Saccharomycetes. On the other hand, Fig. III.17E shows that significant difference in K_S value was observed both in the comparisons of Pezizomycotina genomes and in the comparisons of Saccharomycetes genomes. From these results, regarding Saccharomycetes genomes, higher degree of protein sequence conservation of clustered OGs would be caused by lower substitution rate of coding sequences. Regarding Pezizomycotina genomes, the correlation between protein sequence homology and gene order conservation would be mainly caused by stronger selective pressures to preserve protein sequences, and lower substitution rate of coding sequences also contribute to the correlation.

Based on our results of fungal genome comparisons, the finding of Dandekar *et al.* (1998) that, in prokaryotes, protein sequence of clustered OGs are more conserved than those of isolated OGs could be extended to eukaryotes. This extension would imply the possibility to predict function of eukaryotic genes, or at least fungal genes, based on gene order conservation because the approaches to predicting function of prokaryotic genes are motivated by the finding in Dandekar *et al.* (1998). Since the approaches to predicting gene function based on gene order conservation has been believed to be limitedly useful for prokaryotic genes, further works remain to

determine whether the function of eukaryotic genes can be predicted based on gene order conservation. The correlation between protein sequence homology and gene order conservation is very general trend in prokaryotes because such correlation was observed even between bacterial and archaeal genomes. On the other hand, in fungi, the correlation was observed only in the comparisons of closely related genomes, and was not detected between remotely related genomes. Accordingly, the information of gene order conservation obtained from the comparison of closely related genomes would be more useful to predict function of fungal genes than that obtained from the comparison of remotely related genomes.

5 Conclusion

We proposed a novel workflow that enables sensitive detection of conserved gene clusters by utilizing not only the information of protein sequence similarities but also the information of gene order conservation. Based on the workflow, we confirmed the finding of Dandekar *et al.* (1998) that the degree of protein sequence conservation of clustered OGs is substantially higher than that of isolated OGs in prokaryotes by a large-scale comparison of 101 prokaryotic genomes, and extended to eukaryotes by analyzing 15 fungal genomes. Detailed analyses based on the rate of synonymous substitutions (K_S) and the rate of nonsynonymous substitutions (K_A) unravel that heterogeneous mechanisms would underlie behind the correlation between protein sequence homology and gene order conservation. It is expected that future works will survey whether the finding of Dandekar *et al.* (1998) can be extended to higher eukaryotes, and develop approaches to predicting function of eukaryotic genes based on gene order conservation.

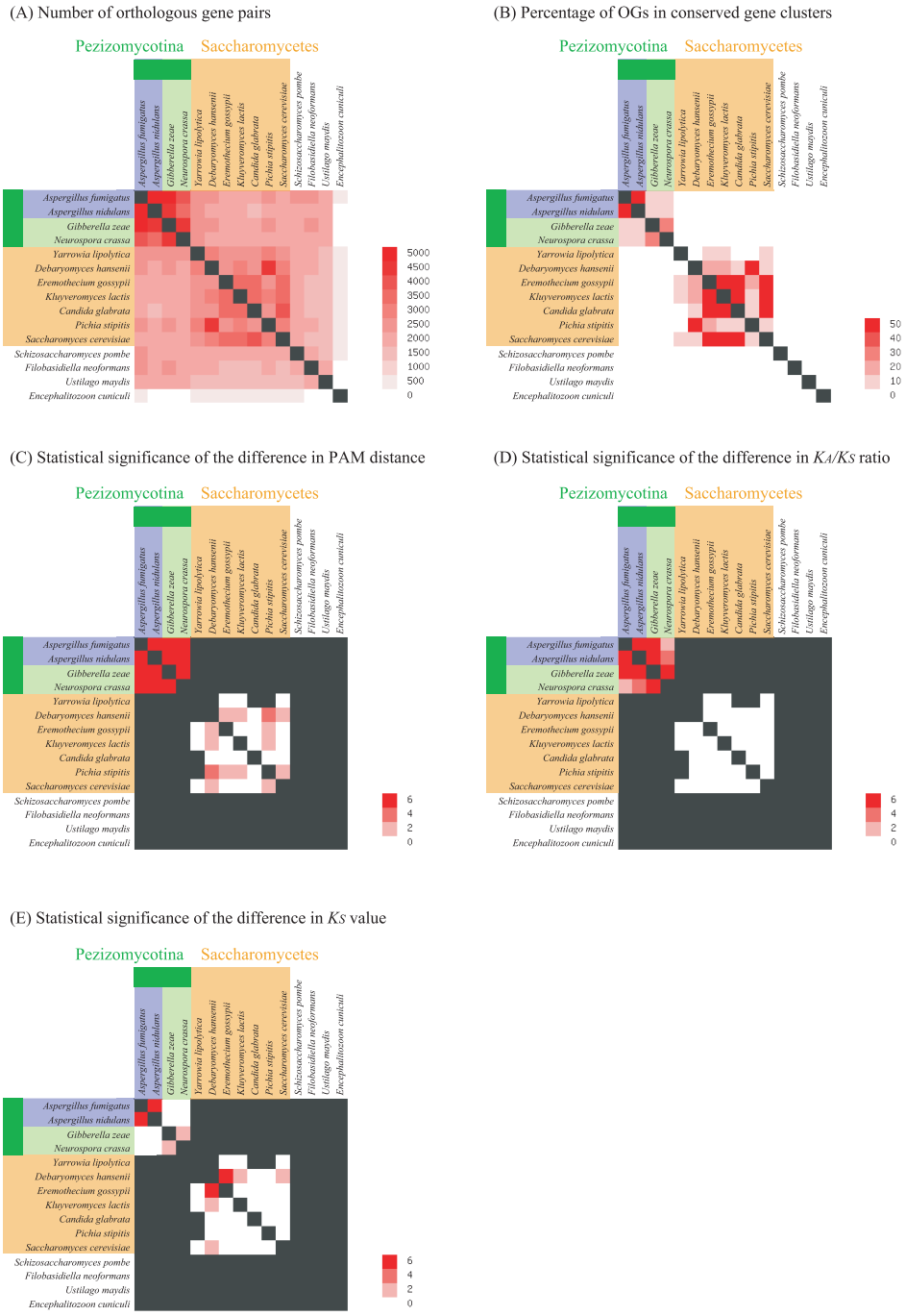


Fig. III.17 Results of comparing fungal genomes. A column or a row in these matrices corresponds to a fungal organism, and the result of comparing two fungal organisms is shown in the corresponding cell. The color of each cell represents the degree of (A) the number of OGs identified by our workflow, (B) the percentage of OGs in conserved gene clusters, (C) the logarithm (base 10) of the p-value of the difference in PAM distance, (D) the logarithm (base 10) of the p-value of the difference in K_A/K_S ratio, and (E) the logarithm (base 10) of the p-value of the difference in K_S value. Analyses corresponding to black-colored cells were not conducted.

Chapter IV

Concluding Remarks

In this dissertation, we embarked on the following theme: inference of evolutionary relationship among biological sequences. Our researches described in Chapters II and III demonstrate that probabilistic models and the decision theory provide powerful frameworks for this direction. In Chapter II, we make use of probabilistic models and the decision theory in order to computationally determine optimal threshold of anchor density. This statistical framework makes it possible to automatically optimize the threshold, and to accurately identify orthologous segments. In Chapter III, we take advantage of probabilistic models and the decision theory so as to integrate heterogeneous information: the information of protein sequence similarity and the information of chromosomal proximity of genes. Thus, a statistical framework that simultaneously takes into account the two types of information is realized, and accurate identification of positional orthologs and sensitive detection of conserved gene clusters are achieved.

The inference of evolutionary relationship among biological sequences is one of the most fundamental problems in comparative genomics; therefore, OSfinder described in Chapter II and OASYS described in Chapter III have much potential to contribute to a wide range of fields in life sciences, including evolutionary biology (Nei and Kumar, 2000; Yang, 2006) and systems biology (Hartwell *et al.*, 1999; Kitano, 2002; Oltvai and Barabási, 2002; Kanehisa *et al.*, 2008). However, since OSfinder and OASYS are still basic softwares in comparative genomics, they are needed to be extended to fit specific purposes in other fields of life sciences in order to make them more beneficial to researchers in those fields. In the remainder of this chapter, further works that would improve the value of our softwares are described from the viewpoints of systems biology and evolutionary biology.

1 Further Works Needed in Systems Biology

In the era of high-throughput sequencing, it is difficult to experimentally determine molecular functions of genes encoded in each newly sequenced genome. Thus, determining orthology relationships among biological sequences and transferring functional annotations from well-studied genomes into newly sequenced genomes are needed (Hulsen *et al.*, 2006; Chen *et al.*, 2006, 2007); the need is especially high for prokaryotic genomes because the number of sequenced prokaryotic genomes have been exponentially increased (Overbeek *et al.*, 2005; Koonin and Wolf, 2008). In order to obtain systems understandings of an organism whose genome has been newly sequenced, knowledge of functions of genes encoded by the genome should be represented by networks (e.g. metabolic and signaling pathways and regulatory networks) (Kanehisa *et al.*, 2006). Thus, accurate computational reconstruction of these networks from genome sequences is the most important challenges to translate the

accumulation of genome sequences into the comprehensive knowledge of biological systems.

The KEGG databases (Kanehisa *et al.*, 2006) provide an effective framework for this direction. The KEGG GENES database employs an original categorization of gene functions, named KEGG Orthology (KO) identifier, to standardize the description of gene functions in the KEGG system (Kanehisa *et al.*, 2006; Moriya *et al.*, 2007). The KO identifiers are assigned to each gene, and their description of gene function is defined based on the KEGG PATHWAY database. The KEGG PATHWAY database provides a reference pathway that contain almost all known biological pathways collected from a number of organisms, and KO identifiers specify which nodes in the reference pathways a gene is mapped onto. Thus, once the KO identifiers are assigned to genes in a newly sequenced genome, organism-specific pathways can be computationally generated by mapping genes in the genome onto the reference pathways based on the KO identifiers (Moriya *et al.*, 2007).

Currently, the KEGG system utilizes KAAS (KEGG Automatic Annotation Server) (Moriya *et al.*, 2007) in order to automatically assign KO identifiers to genes in newly sequenced genomes. The results of KAAS largely depends on the step to identify orthology relationships among genes. KAAS uses similar approaches to reciprocal best hit (RBH) method; therefore, KAAS identifies orthology relationships among genes based only on the information of protein sequence similarity. It has been shown that the information of protein sequence similarity solely is insufficient to accurately determine the orthology relationships (Remm *et al.*, 2001; Mao *et al.*, 2006; Fu *et al.*, 2007; Rödelsperger and Dieterich, 2008), and the information of chromosomal proximity of genes can complement the insufficiency (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999a,b; Snel *et al.*, 2000; Notebaart *et al.*, 2005).

Accordingly, OASYS, which identifies orthology relationships among genes based not only on the information of protein sequence similarity but also the information of chromosomal proximity of genes, would be useful to accurately assign KO identifiers to genes in newly sequenced genomes. In order to demonstrate this expectation, a software that uses OASYS as a core engine to identify orthology relationships and predicts KO identifiers based on the results of OASYS is needed to be developed.

2 Further Works Needed in Evolutionary Biology

Evolutionary processes including nucleotide-level mutations (e.g. base substitutions and short insertions/deletions), gene-level mutations (e.g. lateral gene transfers, gene insertions/deletions, gene duplications, gene fusions, exon shufflings, and intron losses and gains), and chromosome-level mutations (e.g. inversions, transpositions, translocations, chromosomal fusions and fissions, segmental duplications, and large segmental insertions/deletions) are stochastic processes; the occurrence and fixation of these mutations are not deterministic, and are based on certain probability distributions. Thus, probabilistic models can be an effective tool to understand the nature of evolutionary processes. Indeed, the approaches based on probabilistic models have been exten-

sively used to estimate the occurrence probabilities of nucleotide-level mutations (Nei and Kumar, 2000; Holmes, 2005; Yang, 2006; Lunter, 2007; Cartwright, 2009; Heger *et al.*, 2009). These estimates of mutation rates have potential to provide valuable insights into the molecular mechanisms behind nucleotide-level mutations (Pigliucci and Kaplan, 2006).

Although a number of algorithms to estimate the rate of nucleotide-level mutations have been proposed, only a few *ad hoc* algorithms have been proposed to estimate the rate of gene-level and chromosome-level mutations (Pevzner and Tesler, 2003a; Sémon and Wolfe, 2007). It has been shown that, in mammals, the probability to occur genome rearrangements is varied along chromosomes, and extensive reuse of breakpoints from the same short fragile regions have been reported (Armengol *et al.*, 2003; Pevzner and Tesler, 2003b; Bailey *et al.*, 2004). Furthermore, recent researches have revealed that about 70% of genome rearrangements are associated with segmental duplications in the comparison of human and grate apes (Cheng *et al.*, 2005; Kehrer-Sawatzki and Cooper, 2007, 2008), whereas only 40% of genome rearrangements are associated with segmental duplications in the human-gibbon comparison (Bailey and Eichler, 2006; Girirajan *et al.*, 2009). These statistics suggest molecular mechanisms behind genome rearrangements; the occurrence of segmental duplications enhances the the occurrence of genome rearrangements mediated by a molecular mechanism named nonallelic homologous recombination (NAHR) in the evolutionary histories between humans and great apes, whereas microhomology-mediated end-joining (MMEJ) (Yan *et al.*, 2007), fork stalling template switching (FoSTeS) (Lee *et al.*, 2007), or microhomology/microsatellite-induced replication (MMIR) (Payen *et al.*, 2008) would be major molecular mechanisms behind genome rearrangements in the lineage from the common ancestor of primates to gibbons.

As seen in the above example, in order to deepen the understandings of gene- and chromosome-level mutation processes, it is important to develop algorithms to accurately estimate the occurrence probabilities of those mutations, and to unravel correlations between various types of mutations (e.g. correlations between chromosome-level and nucleotide-level mutations). Numerous successes of probabilistic approaches that model nucleotide-level mutation processes suggest that probabilistic approaches that model gene- and chromosome-level mutation processes would be useful for these purposes.

OSfinder is the first algorithm that employs probabilistic models for the problem to identify orthologous segments. As the aim of OSfinder is to accurately identify orthologous segments among multiple genomes, OSfinder uses anchor density as a feature of orthologous segments and the algorithm does not model gene- and chromosome-level mutation processes. Accordingly, although OSfinder provides accurate results of orthologous segments as an input of *ad hoc* algorithms to estimate the occurrence probabilities of gene- and chromosome-level mutations, OSfinder itself can not be applied to estimating those occurrence probabilities in a statistically rigorous manner. Yet, a basic concept of OSfinder (i.e. applying an approach based on probabilistic models to chromosome-level comparisons of genomes) and some mathematical techniques used in OSfinder (e.g. optimization algorithms based on maximum likelihood

approaches) can be a foundation of developing statistical algorithms to model gene- and chromosome-level mutation processes. Indeed, we are now developing such an algorithm based on the basic concept and mathematical techniques with the idea that the problem to identify orthologous segments can be described as an alignment problem that aligns genomes by matching genes; conventional alignment algorithms align genes by matching bases or residues. This idea induces a probabilistic approach that models gene-level mutation processes including gene insertions/deletions and gene duplications. Moreover, the probabilistic approach enables to unravel the correlation between gene-level and nucleotide-level mutations by estimating conditional probabilities to occur nucleotide-level mutations for each state associated with the occurrence of gene-level mutations (details are not shown and will be published elsewhere).

In the era of high-throughput sequencing, genome-level analyses that mine biological insights from rapidly growing repositories of biological sequences are needed, and probabilistic approaches that model gene- and chromosome-level mutation processes would provide effective tools to mine novel insights into evolutionary processes of genomes.

Acknowledgements

I would like to thank Professor Yasubumi Sakakibara who has supervised this study during the period of my bachelor, master and doctor courses. He kindly gave me an opportunity and good environment to complete this thesis. He also taught me the spirit of biological sequence analysis based on probabilistic modeling approaches. Though that, I learned the excitement of developing cutting-edge methodologies in the fields of life science.

I thank all the colleagues of Sakakibara Laboratory in Keio University. Assistant Professor Katsuyuki Yugi gave me advice and comments about the way to write articles. Dr. Kengo Sato kindly taught me the way to implement algorithms for biological sequence analysis as a C program. He also gave me helpful suggestions related to the theories in pattern recognition and machine learning. Dr. Yasunori Osana taught me useful techniques of shell script. With Mr. Kris Popendorf, we discussed how we can contribute to life sciences through computational biology. Mr. Yutaka Saito taught me the experimental methodologies of qRT-PCR. Mr. Souichiro Okuzawa and Mr. Tomoya Tagami were good partners to survey statistical methodologies. Other members of Sakakibara Laboratory provided me with various suggestions and advices. Without their contributions, this thesis wouldn't have been completed.

I also thank Dr. Akihiro Mori, Mr. Masahiro Naruse, and Ms. Emi Niisato for helpful suggestions to improve this thesis. Dr. Akihiro Mori kindly send me his doctoral dissertation. Mr. Masahiro Naruse provided me with useful comments about my presentation of this thesis. Ms. Emi Niisato kindly gave me valuable comments about Chapter I of this thesis.

Lastly, I would like to show my sincere thanks to Professor Yasubumi Sakakibara, Professor Asao Fujiyama, Professor Kotaro Oka, and Associate Professor Nobuhide Doi for examining and judging my doctoral dissertation.

References

H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, Dec 1974.

A. Alexeyenko, I. Tamas, G. Liu, and E. L. Sonnhammer. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22:9–15, Jul 2006.

E. J. Alm, K. H. Huang, M. N. Price, R. P. Koche, K. Keller, I. L. Dubchak, and A. P. Arkin. The MicrobesOnline Web site for comparative genomics. *Genome Res.*, 15:1015–1022, Jul 2005.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, Oct 1990.

S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, Sep 1997.

L. Armengol, M. A. Pujana, J. Cheung, S. W. Scherer, and X. Estivill. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.*, 12:2201–2208, Sep 2003.

C. F. Baer, M. M. Miyamoto, and D. R. Denver. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.*, 8:619–631, Aug 2007.

J. A. Bailey and E. E. Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.*, 7:552–564, Jul 2006.

J. A. Bailey, R. Baertsch, W. J. Kent, D. Haussler, and E. E. Eichler. Hotspots of mammalian chromosomal evolution. *Genome Biol.*, 5:R23, 2004.

S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Res.*, 16:428–435, Mar 2006.

S. M. Barns, C. F. Delwiche, J. D. Palmer, and N. R. Pace. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 93:9188–9193, Aug 1996.

A. K. Baten, S. K. Halgamuge, and B. C. Chang. Fast splice site detection using information content and feature reduction. *BMC Bioinformatics*, 9 Suppl 12:S8, 2008.

J. L. Bennetzen and W. Ramakrishna. Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Mol. Biol.*, 48:821–827, 2002.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Berlin, 2006.

M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, 14:708–715, Apr 2004.

G. Bourque, P. A. Pevzner, and G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.*, 14:507–516, Apr 2004.

G. Bourque, E. M. Zdobnov, P. Bork, P. A. Pevzner, and G. Tesler. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.*, 15:98–110, Jan 2005.

P. P. Calabrese, S. Chakravarty, and T. J. Vision. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, 19 Suppl 1:74–80, 2003.

S. B. Cannon, A. Kozik, B. Chan, R. Michelmore, and N. D. Young. DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol.*, 4:R68, 2003.

R. A. Cartwright. Problems and solutions for estimating indel rates and length distributions. *Mol. Biol. Evol.*, 26:473–480, Feb 2009.

H. Y. Chang. Anatomic demarcation of cells: genes to patterns. *Science*, 326:1206–1207, Nov 2009.

F. Chen, A. J. Mackey, C. J. Stoeckert, and D. S. Roos. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, 34:D363–368, Jan 2006.

F. Chen, A. J. Mackey, J. K. Vermunt, and D. S. Roos. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2:e383, 2007.

Z. Cheng, M. Ventura, X. She, P. Khaitovich, T. Graves, K. Osoegawa, D. Church, P. DeJong, R. K.

Wilson, S. Pääbo, M. Rocchi, and E. E. Eichler. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, 437:88–93, Sep 2005.

G. Cochrane, R. Akhtar, J. Bonfield, L. Bower, F. Demiralp, N. Faruque, R. Gibson, G. Hoad, T. Hubbard, C. Hunter, M. Jang, S. Juhos, R. Leinonen, S. Leonard, Q. Lin, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, S. Plaister, R. Radhakrishnan, S. Robinson, S. Sobhany, P. T. Hoopen, R. Vaughan, V. Zalunin, and E. Birney. Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res.*, 37:19–25, Jan 2009.

T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, 23:324–328, Sep 1998.

P. S. Dehal and J. L. Boore. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics*, 7:201, 2006.

A. L. Delcher, K. A. Bratke, E. C. Powers, and S. L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23:673–679, Mar 2007.

C. N. Dewey. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol. Biol.*, 395:221–236, 2007.

C. N. Dewey, P. M. Huggins, K. Woods, B. Sturmfels, and L. Pachter. Parametric alignment of Drosophila genomes. *PLoS Comput. Biol.*, 2:e73, Jun 2006.

R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, Cambridge, 1998.

J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author, Department of Genome Sciences, University of Washington, Seattle, USA*, 2005.

R. A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930.

W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst. Zool.*, 19:99–113, Jun 1970.

R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Computer J.*, 7: 149–154, 1964.

A. K. Fox, B. B. Tuch, and J. H. Chuang. Measuring the prevalence of regional mutation rates: an analysis of silent substitutions in mammals, fungi, and insects. *BMC Evol. Biol.*, 8:186, 2008.

K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, 32:W273–279, Jul 2004.

Z. Fu, X. Chen, V. Vacic, P. Nan, Y. Zhong, and T. Jiang. MSOAR: a high-throughput ortholog assignment system based on genome rearrangement. *J. Comput. Biol.*, 14:1160–1175, Nov 2007.

R. A. Gibbs, G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, S. Scherer, G. Scott, D. Steffen, K. C. Worley, P. E. Burch, G. Okwuonu, S. Hines, L. Lewis, C. DeRamo, O. Delgado, S. Dugan-Rocha, G. Miner, M. Morgan, A. Hawes, R. Gill, Celera, R. A. Holt, M. D. Adams, P. G. Amanatides, H. Baden-Tillson, M. Barnstead, S. Chin, C. A. Evans, S. Ferriera, C. Fosler, A. Glodek, Z. Gu, D. Jennings, C. L. Kraft, T. Nguyen, C. M. Pfannkoch, C. Sitter, G. G. Sutton, J. C. Venter, T. Woodage, D. Smith, H. M. Lee, E. Gustafson, P. Cahill, A. Kana, L. Doucette-Stamm, K. Weinstock, K. Fechtel, R. B. Weiss, D. M. Dunn, E. D. Green, R. W. Blakesley, G. G. Bouffard, P. J. De Jong, K. Osoegawa, B. Zhu, M. Marra, J. Schein, I. Bosdet, C. Fjell, S. Jones, M. Krzywinski, C. Mathewson, A. Siddiqui, N. Wye, J. McPherson, S. Zhao, C. M. Fraser, J. Shetty, S. Shatsman, K. Geer, Y. Chen, S. Abramzon, W. C. Nierman, P. H. Havlak, R. Chen, K. J. Durbin, A. Egan, Y. Ren, X. Z. Song, B. Li, Y. Liu, X. Qin, S. Cawley, K. C. Worley, A. J. Cooney, L. M. D'Souza, K. Martin, J. Q. Wu, M. L. Gonzalez-Garay, A. R. Jackson, K. J. Kalafus, M. P. McLeod, A. Milosavljevic, D. Virk, A. Volkov, D. A. Wheeler, Z. Zhang, J. A. Bailey, E. E. Eichler, E. Tuzun, E. Birney, E. Mongin, A. Ureta-Vidal, C. Woodwark, E. Zdobnov, P. Bork, M. Suyama, D. Torrents, M. Alexandersson, B. J. Trask, J. M. Young, H. Huang, H. Wang, H. Xing, S. Daniels, D. Gietzen, J. Schmidt, K. Stevens, U. Vitt, J. Wingrove, F. Camara, M. Mar Alba, J. F. Abril, R. Guigo, A. Smit, I. Dubchak, E. M. Rubin, O. Couronne, A. Poliakov, N. Hubner, D. Ganten, C. Goesele, O. Hummel, T. Kreitler, Y. A. Lee, J. Monti, H. Schulz, H. Zimdahl, H. Himmelbauer, H. Lehrach, H. J. Jacob, S. Bromberg, J. Gullings-Handley, M. I. Jensen-Seaman, A. E. Kwitek, J. Lazar, D. Pasko, P. J. Tonellato, S. Twigger, C. P. Ponting, J. M. Duarte, S. Rice, L. Goodstadt, S. A. Beatson, R. D. Emes, E. E. Winter, C. Webber, P. Brandt, G. Nyakatura, M. Adetobi, F. Chiaromonte, L. Elnitski, P. Eswara, R. C. Hardison, M. Hou, D. Kolbe, K. Makova, W. Miller, A. Nekrutenko, C. Riemer, S. Schwartz, J. Taylor, S. Yang, Y. Zhang, K. Lindpaintner, T. D. Andrews, M. Caccamo, M. Clamp, L. Clarke, V. Curwen,

R. Durbin, E. Eyras, S. M. Searle, G. M. Cooper, S. Batzoglou, M. Brudno, A. Sidow, E. A. Stone, J. C. Venter, B. A. Payseur, G. Bourque, C. Lopez-Otin, X. S. Puente, K. Chakrabarti, S. Chatterji, C. Dewey, L. Pachter, N. Bray, V. B. Yap, A. Caspi, G. Tesler, P. A. Pevzner, D. Haussler, K. M. Roskin, R. Baertsch, H. Clawson, T. S. Furey, A. S. Hinrichs, D. Karolchik, W. J. Kent, K. R. Rosenbloom, H. Trumbower, M. Weirauch, D. N. Cooper, P. D. Stenson, B. Ma, M. Brent, M. Arumugam, D. Shteynberg, R. R. Copley, M. S. Taylor, H. Riethman, U. Mudunuri, J. Peterson, M. Guyer, A. Felsenfeld, S. Old, S. Mockrin, and F. Collins. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521, Apr 2004.

S. Girirajan, L. Chen, T. Graves, T. Marques-Bonet, M. Ventura, C. Fronick, L. Fulton, M. Rocchi, R. S. Fulton, R. K. Wilson, E. R. Mardis, and E. E. Eichler. Sequencing human-gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites. *Genome Res.*, 19:178–190, Feb 2009.

B. J. Haas, A. L. Delcher, J. R. Wortman, and S. L. Salzberg. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, 20:3643–3646, Dec 2004.

T. Hachiya and Y. Sakakibara. Sensitive detection of conserved gene clusters unravels the evolutionary forces behind the correlation between protein sequence homology and gene order conservation. *Genes, Genomes and Genomics*, 3:31–45, 2009.

T. Hachiya, Y. Osana, K. Popendorf, and Y. Sakakibara. Accurate identification of orthologous segments among multiple genomes. *Bioinformatics*, 25:853–860, Apr 2009.

S. Hampson, A. McLysaght, B. Gaut, and P. Baldi. LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res.*, 13:999–1010, May 2003.

M. V. Han, J. P. Demuth, C. L. McGrath, C. Casola, and M. W. Hahn. Adaptive evolution of young gene duplicates in mammals. *Genome Res.*, 19:859–867, May 2009.

R. C. Hardison, K. M. Roskin, S. Yang, M. Diekhans, W. J. Kent, R. Weber, L. Elnitski, J. Li, M. O'Connor, D. Kolbe, S. Schwartz, T. S. Furey, S. Whelan, N. Goldman, A. Smit, W. Miller, F. Chiaromonte, and D. Haussler. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.*, 13:13–26, Jan 2003.

L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:47–52, Dec 1999.

S. B. Hedges and P. Shah. Comparison of mode estimation methods and application in molecular clock analysis. *BMC Bioinformatics*, 4:31, Jul 2003.

A. Heger, C. P. Ponting, and I. Holmes. Accurate estimation of gene evolutionary rates using XRATE, with an application to transmembrane proteins. *Mol. Biol. Evol.*, 26:1715–1721, Aug 2009.

D. P. Herlemann, O. Geissinger, W. Ikeda-Ohtsubo, V. Kunin, H. Sun, A. Lapidus, P. Hugenholtz, and A. Brune. Genomic analysis of "Elusimicrobium minutum," the first cultivated representative of the phylum "Elusimicrobia" (formerly termite group 1). *Appl. Environ. Microbiol.*, 75:2841–2849, May 2009.

L. W. Hillier, R. D. Miller, S. E. Baird, A. Chinwalla, L. A. Fulton, D. C. Koboldt, and R. H. Waterston. Comparison of C. elegans and C. briggsae genome sequences reveals extensive conservation of chromosome organization and synteny. *PLoS Biol.*, 5:e167, Jul 2007.

A. E. Hirsh and H. B. Fraser. Protein dispensability and rate of evolution. *Nature*, 411:1046–1049, Jun 2001.

I. Holmes. Using evolutionary Expectation Maximization to estimate indel rates. *Bioinformatics*, 21:2294–2300, May 2005.

M. Hryniewicz, A. Sirko, A. Palucha, A. Böck, and D. Hulanicka. Sulfate and thiosulfate transport in Escherichia coli K-12: identification of a gene encoding a novel protein involved in thiosulfate binding. *J. Bacteriol.*, 172:3358–3366, Jun 1990.

T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-

Vidal, J. Vogel, S. White, C. Woodwark, and E. Birney. Ensembl 2005. *Nucleic Acids Res.*, 33: D447–453, Jan 2005.

T. J. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res.*, 35:D610–617, Jan 2007.

H. Huber, M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, and K. O. Stetter. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, 417:63–67, May 2002.

T. Hulsen, J. de Vlieg, and P. M. Groenen. PhyloPat: phylogenetic pattern analysis of eukaryotic genes. *BMC Bioinformatics*, 7:398, 2006.

M. Huynen, B. Snel, W. Lathe, and P. Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, 10:1204–1210, Aug 2000.

Z. Jiang, R. Hubley, A. Smit, and E. E. Eichler. DupMasker: a tool for annotating primate segmental duplications. *Genome Res.*, 18:1362–1368, Aug 2008.

I. K. Jordan, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, 12:962–968, Jun 2002.

M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, 34:D354–357, Jan 2006.

M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, 36:D480–484, Jan 2008.

D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, 36:D773–779, Jan 2008.

H. Kehrer-Sawatzki and D. N. Cooper. Structural divergence between the human and chimpanzee genomes. *Hum. Genet.*, 120:759–778, Feb 2007.

H. Kehrer-Sawatzki and D. N. Cooper. Molecular mechanisms of chromosomal rearrangement during primate evolution. *Chromosome Res.*, 16:41–56, 2008.

W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.*, 100:11484–11489, Sep 2003.

H. Kitano. Computational systems biology. *Nature*, 420:206–210, Nov 2002.

E. V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39:309–338, 2005.

E. V. Koonin. Evolution of genome architecture. *Int. J. Biochem. Cell Biol.*, 41:298–306, Feb 2009.

E. V. Koonin and Y. I. Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, 36:6688–6719, Dec 2008.

E. V. Koonin, A. R. Mushegian, and K. E. Rudd. Sequencing and analysis of bacterial genomes. *Curr. Biol.*, 6:404–416, Apr 1996.

C. Kosiol and N. Goldman. Different versions of the Dayhoff rate matrix. *Mol. Biol. Evol.*, 22: 193–199, Feb 2005.

R. M. Kuhn, D. Karolchik, A. S. Zweig, H. Trumbower, D. J. Thomas, A. Thakkapallayil, C. W. Sugnet, M. Stanke, K. E. Smith, A. Siepel, K. R. Rosenbloom, B. Rhead, B. J. Raney, A. Pohl, J. S. Pedersen, F. Hsu, A. S. Hinrichs, R. A. Harte, M. Diekhans, H. Clawson, G. Bejerano, G. P. Barber, R. Baertsch, D. Haussler, and W. J. Kent. The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, 35:D668–673, Jan 2007.

J. A. Lee, C. M. Carvalho, and J. R. Lupski. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, 131:1235–1247, Dec 2007.

J. Y. Lee and A. K. Nandi. Maximum likelihood parameter estimation of the asymmetric generalised gaussian family of distributions. *Proc. SPW-HOS*, pages 255–258, Jun 1999.
F. Lemoine, O. Lespinet, and B. Labedan. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC Evol. Biol.*, 7:237, 2007.
J. Li, S. K. Halgamuge, C. I. Kells, and S. L. Tang. Gene function prediction based on genomic context clustering and discriminative learning: an application to bacteriophages. *BMC Bioinformatics*, 8 Suppl 4:S6, 2007.
L. Li, C. J. Stoeckert, and D. S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13:2178–2189, Sep 2003.
K. Liolios, K. Mavromatis, N. Tavernarakis, and N. C. Kyrpides. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, 36:D475–479, Jan 2008.
T. M. Lowe and S. R. Eddy. A computational screen for methylation guide snoRNAs in yeast. *Science*, 283:1168–1171, Feb 1999.
G. Lunter. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, 23:i289–296, Jul 2007.
G. Lunter. Dog as an outgroup to human and mouse. *PLoS Comput. Biol.*, 3:e74, Apr 2007a.
J. Ma, L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans, W. J. Kent, M. Blanchette, D. Haussler, and W. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Res.*, 16: 1557–1565, Dec 2006.
D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.
H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, 18:50–60, 1947.
F. Mao, Z. Su, V. Olman, P. Dam, Z. Liu, and Y. Xu. Mapping of orthologous genes in the context of biological pathways: An application of integer programming. *Proc. Natl. Acad. Sci. U.S.A.*, 103:129–134, Jan 2006.
E. H. Margulies, G. M. Cooper, G. Asimenos, D. J. Thomas, C. N. Dewey, A. Siepel, E. Birney, D. Keefe, A. S. Schwartz, M. Hou, J. Taylor, S. Nikolaev, J. I. Montoya-Burgos, A. Löytynoja, S. Whelan, F. Pardi, T. Massingham, J. B. Brown, P. Bickel, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, E. A. Stone, K. R. Rosenbloom, W. J. Kent, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, A. Hinrichs, H. Trumbower, H. Clawson, A. Zweig, R. M. Kuhn, G. Barber, R. Harte, D. Karolchik, M. A. Field, R. A. Moore, C. A. Matthewson, J. E. Schein, M. A. Marra, S. E. Antonarakis, S. Batzoglou, N. Goldman, R. Hardison, D. Haussler, W. Miller, L. Pachter, E. D. Green, and A. Sidow. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.*, 17:760–774, Jun 2007.
J. H. Miller and J. B. Thomas. Detectors for discrete-time signals in non-Gaussian noise. *IEEE Transaction on Information Theory*, 18:241–250, Mar 1972.
R. C. Moore and M. D. Purugganan. The early stages of duplicate gene evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 100:15682–15687, Dec 2003.
Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, 35:W182–185, Jul 2007.
W. J. Murphy, D. M. Larkin, A. Everts-van der Wind, G. Bourque, G. Tesler, L. Auvil, J. E. Beever, B. P. Chowdhary, F. Galibert, L. Gatzke, C. Hitte, S. N. Meyers, D. Milan, E. A. Ostrander, G. Pape, H. G. Parker, T. Raudsepp, M. B. Rogatcheva, L. B. Schook, L. C. Skow, M. Welge, J. E. Womack, S. J. O'brien, P. A. Pevzner, and H. A. Lewin. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309:613–617, Jul 2005.
A. R. Mushegian and E. V. Koonin. Gene order is not conserved in bacterial evolution. *Trends Genet.*, 12:289–290, Aug 1996.
M. Nei and S. Kumar. *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford, 2000.

S. Nikolaev, J. I. Montoya-Burgos, E. H. Margulies, J. Rougemont, B. Nyffeler, and S. E. Antonarakis. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. *PLoS Genet.*, 3:e2, Jan 2007.

R. A. Notebaart, M. A. Huynen, B. Teusink, R. J. Siezen, and B. Snel. Correlation between sequence conservation and the genomic context after gene duplication. *Nucleic Acids Res.*, 33:6164–6171, 2005.

S. Ohno. *Evolution by Gene Duplication*. Springer-Verlag, Berlin, 1970.

Z. N. Oltvai and A. L. Barabási. Systems biology. Life's complexity pyramid. *Science*, 298:763–764, Oct 2002.

R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol. (Gedrukt)*, 1:93–108, 1999a.

R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.*, 96:2896–2901, Mar 1999b.

R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, 33:5691–5702, 2005.

C. Payen, R. Koszul, B. Dujon, and G. Fischer. Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. *PLoS Genet.*, 4: e1000175, 2008.

F. Pazos and A. Valencia. Protein co-evolution, co-adaptation and interactions. *EMBO J.*, 27: 2648–2655, Oct 2008.

P. Pevzner and G. Tesler. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.*, 13:37–45, Jan 2003a.

P. Pevzner and G. Tesler. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 100:7672–7677, Jun 2003b.

M. Pigliucci and J. Kaplan. *The Conceptual Foundations of Evolutionary Biology*. The University of Chicago Press, Chicago, 2006.

K. Popendorf, Y. Osana, T. Hachiya, and Y. Sakakibara. Murasaki – homology detection across multiple large-scale genomes. *Fifth Annual RECOMB Satellite Workshop on Comparative Genomics, San Diego, USA*, Sep 2007.

M. N. Price, K. H. Huang, E. J. Alm, and A. P. Arkin. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, 33:880–892, 2005.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 35: D61–65, Jan 2007.

M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, 314:1041–1052, Dec 2001.

P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16:276–277, Jun 2000.

M. C. Rivera, R. Jain, J. E. Moore, and J. A. Lake. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U.S.A.*, 95:6239–6244, May 1998.

C. Rödelsperger and C. Dieterich. Syntenator: multiple gene order alignments with a gene-specific scoring function. *Algorithms Mol Biol*, 3:14, 2008.

F. Rodriguez-Trelles, R. Tarrío, and F. J. Ayala. Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine dehydrogenase gene. *Proc. Natl. Acad. Sci. U.S.A.*, 100:13413–13417, Nov 2003.

S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with BLASTZ. *Genome Res.*, 13:103–107, Jan 2003.

M. Sémon and K. H. Wolfe. Rearrangement rate following the whole-genome duplication in teleosts. *Mol. Biol. Evol.*, 24:860–867, Mar 2007.

A. U. Sinha and J. Meller. Cinteny: flexible analysis and visualization of synteny and genome

rearrangements in multiple organisms. *BMC Bioinformatics*, 8:82, 2007.

A. Sirko, M. Hryniewicz, D. Hulanicka, and A. Böck. Sulfate and thiosulfate transport in Escherichia coli K-12: nucleotide sequence and expression of the cysTWAM gene cluster. *J. Bacteriol.*, 172: 3351–3357, Jun 1990.

B. Snel, G. Lehmann, P. Bork, and M. A. Huynen. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.*, 28:3442–3444, Sep 2000.

C. Soderlund, W. Nelson, A. Shoemaker, and A. Paterson. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Res.*, 16:1159–1168, Sep 2006.

R. Song, V. Llaca, and J. Messing. Mosaic organization of orthologous sequences in grass genomes. *Genome Res.*, 12:1549–1555, Oct 2002.

J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, 12:1611–1618, Oct 2002.

R. L. Tatusov, A. R. Mushegian, P. Bork, N. P. Brown, W. S. Hayes, M. Borodovsky, K. E. Rudd, and E. V. Koonin. Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. *Curr. Biol.*, 6:279–291, Mar 1996.

R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278:631–637, Oct 1997.

R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003.

A. Tesei and C. S. Regazzoni. HOS-based generalized noise pdf models for signal detection optimization. *Signal Processing*, 65:267–281, Mar 1998.

G. Tesler. GRIMM: genome rearrangements web server. *Bioinformatics*, 18:492–493, Mar 2002.

K. Thornton and M. Long. Excess of amino acid substitutions relative to polymorphism between X-linked duplications in Drosophila melanogaster. *Mol. Biol. Evol.*, 22:273–284, Feb 2005.

E. J. Vallender, J. E. Paschall, C. M. Malcom, B. T. Lahn, and G. J. Wyckoff. SPEED: a molecular-evolution-based database of mammalian orthologous groups. *Bioinformatics*, 22:2835–2837, Nov 2006.

K. Vandepoele, Y. Saeys, C. Simillion, J. Raes, and Y. Van De Peer. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Res.*, 12:1792–1801, Nov 2002.

A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19: 327–335, Feb 2009.

D. P. Wall, H. B. Fraser, and A. E. Hirsh. Detecting putative orthologs. *Bioinformatics*, 19:1710–1711, Sep 2003.

B. Wallace. Coadaptation revisited. *J. Hered.*, 82:89–95, 1991.

P. P. Wang and I. Ruvinsky. Computational prediction of Caenorhabditis box H/ACA snoRNAs using genomic properties of their host genes. *RNA*, Dec 2009.

H. Watanabe, H. Mori, T. Itoh, and T. Gojobori. Genome plasticity as a paradigm of eubacteria evolution. *J. Mol. Evol.*, 44 Suppl 1:57–64, 1997.

R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigó, M. Guyer, R. C. Hardison,

D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S. P. Yang, E. M. Zdobnov, M. C. Zody, and E. S. Lander. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, Dec 2002.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

Y. I. Wolf, I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.*, 11:356–372, Mar 2001.

K. H. Wolfe, P. M. Sharp, and W. H. Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337:283–285, Jan 1989.

C. T. Yan, C. Boboila, E. K. Souza, S. Franco, T. R. Hickernell, M. Murphy, S. Gumaste, M. Geyer, A. A. Zarrin, J. P. Manis, K. Rajewsky, and F. W. Alt. IgH class switching and translocations use a robust non-classical end-joining pathway. *Nature*, 449:478–482, Sep 2007.

Z. Yang. *Computational Molecular Evolution*. Oxford University Press, Oxford, 2006.

Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, 13:555–556, Oct 1997.

Z. Yang and R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, 17:32–43, Jan 2000.

Z. Yang, R. Nielsen, N. Goldman, and A. M. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449, May 2000.

H. Zhang, Y. Sekiguchi, S. Hanada, P. Hugenholtz, H. Kim, Y. Kamagata, and K. Nakamura. Gemmatimonas aurantiaca gen. nov., sp. nov., a gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov. *Int. J. Syst. Evol. Microbiol.*, 53:1155–1163, Jul 2003.

J. Zhang, H. F. Rosenberg, and M. Nei. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. U.S.A.*, 95:3708–3713, Mar 1998.

X. H. Zheng, F. Lu, Z. Y. Wang, F. Zhong, J. Hoover, and R. Mural. Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics*, 21:703–710, Mar 2005.

Appendix A

Software Web Sites

OSfinder – Orthologous Segment finder

The OSfinder software implemented as a C++ program is freely available at the following URL under the GNU General Public License.

http://osfinder.dna.bio.keio.ac.jp

OASYS – Ortholog Assignment based on SYnteny and Sequence information

The OASYS software implemented as a C++ program is freely available at the following URL under the GNU General Public License.

http://oasys.dna.bio.keio.ac.jp

Appendix B

List of Publications

Journal Papers

1. Hachiya, T., Osana Y., Popendorf, K. and Sakakibara, Y. Accurate identification of orthologous segments among multiple genomes. *Bioinformatics* **25**, 853–860 (2009).
   - The research described in Chapter II was reported in this paper.
2. Hachiya, T. and Sakakibara, Y. Sensitive detection of conserved gene clusters unravels the evolutionary forces behind the correlation between protein sequence homology and gene order conservation. *Genes, Genomes and Genomics* **3**, 31–45 (2009).
   - A part of the research described in Chapter III was reported in this paper.

Conference Proceedings (peer-reviewed full-length papers)

1. Hachiya, T. and Sakakibara, Y. Searching biologically plausible synteny blocks among multiple genomes. *Proceedings of the 2005 International Joint Conference of InCoB, AASBi and KSBI*, 113–117, Busan, Korea, September (2005).

International Conferences (poster presentation)

1. Hachiya, T. and Sakakibara, Y. Stochastic local genome alignment and comprehensive search for conserved gene clusters. *The 17th International Conference on Genome Informatics*, Yokohama, Japan, December (2006)
2. Popendorf, K., Osana, Y., Hachiya, T. and Sakakibara, Y. Murasaki – homology detection across multiple large-scale genomes. *The 5th Annual RECOMB Satellite Workshop on Comparative Genomics*, San Diego, USA, September (2007)
3. Hachiya, T., Osana, Y., Popendorf, K. and Sakakibara, Y. Rearrangement events have synchronized with nucleotide substitutions during mammalian evolutionary history. *The 21st International Mammalian Genome Conference*, Kyoto, Japan, October (2007).
4. Hachiya, T., Osana, Y., Popendorf, K. and Sakakibara, Y. OSfinder: A tool for accurate orthology mapping and its application to mammalian genomes. *The 7th International Workshop on Advanced Genomics*, Tokyo, Japan, November (2007).
5. Kawarama, J., Hase, S., Hachiya, T., Hotta, K. and Sakakibara, Y. Genome-wide detections of non-coding RNAs on *Ciona intestinalis* genome: from *in silico* search of snoRNA to full-length sequencing and expression analysis. *The 5th International Tunicate Meeting*, Okinawa, Japan, June (2009).