BIOINFORMATICS ORIGINAL PAPER
Vol. 26 no. 17 2010, pages 2076–2084
doi:10.1093/bioinformatics/btq339

Genome analysis

Advance Access publication June 27, 2010

EGM: encapsulated gene-by-gene matching to identify gene orthologs and homologous segments in genomes

Khalid Mahmood[1,2], Arun S. Konagurthu[3,4], Jiangning Song[1], Ashley M. Buckle[1], Geoffrey I. Webb[5,*] and James C. Whisstock[1,2,*]

[1]Department of Biochemistry and Molecular Biology, [2]ARC Centre of Excellence in Structural and Functional Microbial Genomics, Monash University, Clayton, VIC 3800, [3]National ICT Australia Victoria Research Laboratory, Department of Electrical and Electronic Engineering, [4]Department of Computer Science and Software Engineering, University of Melbourne, Parkville, VIC 3010 and [5]Clayton School of Information Technology, Monash University, Clayton, VIC 3800, Australia

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Identification of functionally equivalent genes in different species is essential to understand the evolution of biological pathways and processes. At the same time, identification of strings of conserved orthologous genes helps identify complex genomic rearrangements across different organisms. Such an insight is particularly useful, for example, in the transfer of experimental results between different experimental systems such as Drosophila and mammals.

Results: Here, we describe the Encapsulated Gene-by-gene Matching (EGM) approach, a method that employs a graph matching strategy to identify gene orthologs and conserved gene segments. Given a pair of genomes, EGM constructs a global gene match for all genes taking into account gene context and family information. The Hungarian method for identifying the maximum weight matching in bipartite graphs is employed, where the resulting matching reveals one-to-one correspondences between nodes (genes) in a manner that maximizes the gene similarity and context.

Conclusion: We tested our approach by performing several comparisons including a detailed Human versus Mouse genome mapping. We find that the algorithm is robust and sensitive in detecting orthologs and conserved gene segments. EGM can sensitively detect rearrangements within large and small chromosomal segments. The EGM tool is fully automated and easy to use compared to other more complex methods that also require extensive manual intervention and input.

Availability: The EGM software, Supplementary information and other tools are available online from http://vbc.med.monash.edu.au/~kmahmood/EGM

Contacts: james.whisstock@monash.edu; geoff.webb@monash.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 26, 2010; revised and accepted on June 18, 2010

1 INTRODUCTION

Inferring orthology relationships and identifying conserved gene segments are essential tasks in genome comparison. Identification of matching genes and their contiguous strings provide a perspective on how genomes are related, how they function and how the species evolve. Segments of conserved homologous genes across species are termed as *conserved synteny* (Dewey *et al.*, 2006). A core task in identifying conserved synteny is the assignment of putative orthologous genes. The presence of several homologous genes across the species makes this a complex task. However, global genome properties used in comparative studies have shown that strings of genes are well preserved, especially in closely related organisms (Blanchette *et al.*, 2009; Sankoff, 1999; Swidan *et al.*, 2006). Further, genes with evolutionary relationship tend to conserve their genomic context, especially in closely related species (Ayala, 1994; Blanchette, 2007; Catchen *et al.*, 2009; Dandekar *et al.*, 1998; Lathe *et al.*, 2000; Rogozin *et al.*, 2004; Sankoff, 1999; Tamames, 2001; Tamames *et al.*, 1997). Therefore, gene context and neighborhood provides important information for assigning the best orthologs.

Commonly, methods used to identify orthologs and synteny between a pair of genomes work by identifying homologous genes, followed by their expansion to build larger collinear blocks of similar genes (Peng *et al.*, 2009). These techniques can be grouped into two categories from an algorithmic perspective (Abouelhoda, 2005; Chain *et al.*, 2003): (i) the first category is similar to an 'alignment' of molecular sequences. Semi-automated techniques have been used to align genome sequences mainly when information regarding the conserved regions is known, such as, in the comparison of *Mycoplasma genitalium* and *M.pneumonioae* species presented in Himmelreich *et al.* (1997). Often, these semi-automated techniques work by initially identifying conserved regions using all-against-all gene comparisons. Manual identification of conserved regions is challenging, especially when larger genomes with complex rearrangements are involved (Goldberg *et al.*, 2000). These regions are then separated and aligned using standard tools such as BLAST (Altschul *et al.*, 1990, 1997). In addition, automated methods to determine conserved regions based on the comparison of fixed length segments (string of genes) have been developed, such as those used to compare the genomes of *Caenorhabditis briggsae* and *C.elegans* in Kent *et al.* (2000). However, they cannot be efficiently applied in the case of large sequences (Peng *et al.*, 2006; Pevzner and Tesler, 2003). (ii) The second classification of techniques can be categorized as 'matching' of corresponding homologous genes. Unlike alignments, where the precedence of the nodes (genes) is maintained, matching-based techniques work

*To whom correspondence should be addressed.

by identifying small segments of highly similar genes that are grouped to produce a comparative map containing gene matches. Tools such as BLAST are often used, as in Waterston *et al.* (2002), for obtaining all-versus-all comparisons between genes to calculate initial gene correspondences without direct knowledge of their genomic positions and context. More sophisticated approaches include those that either group together traditional molecular alignments (Kent *et al.*, 2003), or conserved gene segments include: Pash (Kalafus *et al.*, 2004), MUMmer (Kurtz *et al.*, 2004), ABWGAT (Das *et al.*, 2009), ADHoRe (Vandepoele *et al.*, 2002), FISH (Calabrese *et al.*, 2003), CHSMiner (Wang *et al.*, 2009), DAGchainer (Haas *et al.*, 2004) and ABS (Peng *et al.*, 2009). Approaches modeled on the bipartite graph-matching problem have also been developed for comparing genomes: AuberGene (Szklarczyk and Heringa, 2006), MSOAR (Fu *et al.*, 2007; Shi *et al.*, 2010) and others like Bansal *et al.* (1998) and Kellis *et al.* (2004). Further, see Salse *et al.* (2009); Sankoff, (1999); Sankoff and Nadeau, (2000) for some of the important earlier and more recent works and applications. More recently, statistical and machine learning strategies have also been successfully employed to improve the accuracy in ortholog mapping, such as OSfinder (Hachiya *et al.*, 2009). However, there is an increasing recognition of the importance of detecting conserved gene context and gene neighbors in a global fashion as this can provide valuable information for inferring function and evolutionary relationships (Huynen *et al.*, 2000a,b; Lemoine *et al.*, 2007; Nadeau and Taylor, 1984; Swidan *et al.*, 2006).

In this work, we report a fully automated method Encapsulated Gene-by-gene Matching (EGM) that automatically identifies gene orthologs and homologous non-linear strings of genes between a pair of genomes. The approach is able to reveal conserved gene segments or syntenic regions, best gene orthologs, evolutionary rearrangements and presents the results in a general comparative map. EGM relies on the sequence similarity, gene context and orientation information to identify the best one-to-one gene matches or putative orthologs. This is useful when comparing evolutionary conservation of the gene order in addition to examining the genome construction in terms of protein families to help understand and predict gene function (Bandyopadhyay *et al.*, 2006; Goldberg *et al.*, 2000; Rasmussen and Kellis, 2007; Sharan *et al.*, 2005; Shi *et al.*, 2010; Wu *et al.*, 2006). We use the notations described by Sankoff (1999), where genomes are represented as set of sequences in which individual proteins are represented by their corresponding family identifiers. To resolve gene correspondence between a pair of genomes, an analogy is drawn between the gene-matching problem and the *Linear Assignment* problem. Essentially, the task of identifying gene orthologs is transformed to a *graph assignment* problem where the goal is to match genes between species such that the sum total similarity score of all matched genes is maximized. For this purpose, a matched segment length-dependent edge-weighting scheme is used to calculate matching in a weighted bipartite graph. The Hungarian method (Kuhn, 1955) for the graph assignment is employed to identify the best one-to-one orthologs that maximizes conserved synteny, while filtering spurious matches. We performed several experiments to evaluate the performance of EGM and compared with other popular methods. Our results clearly show that EGM is effective in identifying best one-to-one gene matches (orthologous pairs) in a straightforward manner.

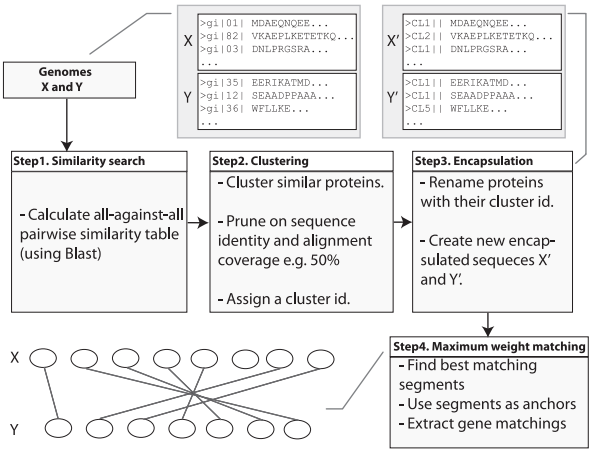

Fig. 1. The input is two FASTA formatted files containing protein sequences in their chromosomal order. Initially, all possible gene-by-gene associations are identified (step 1). This information is then used to calculate PHAMs or clusters of proteins within the two species (step 2). Each protein in the two species is then relabeled using a cluster identifier to give the *encapsulated* form of the genomes (step 3). Next the problem of identifying the best bijective match for each gene is solved using the Hungarian method (step 4). The result is a one-to-one gene map between the two genomes; clearly showing homologous conserved gene segments, their context, as well as their complex evolutionary rearrangement events.

2 METHODS

2.1 EGM overview

We describe here the EGM algorithm that addresses the whole genome gene-matching problem. (For overview see Fig. 1.) The pipeline first constructs a similarity matrix showing amino acid sequence similarity for all protein pairs within and across the two species. Next, to add the notion of protein families, we construct clusters of similar proteins across the two species. We relabel each protein sequence in the genome data using an encapsulation method, characterizing each genome as a sequence of cluster identifiers. The comparison between two encapsulated genomes is modeled as a bipartite graph with weighted edges. Each node in the graph is a protein sequence (represented by its encapsulated cluster label). There is an edge between two nodes (one from each genome) when the corresponding sequences are similar. An *ad-hoc* edge weight matrix is computed, where the edge weights are reinforced when contiguous stretches of nodes in the two genomes share a strong sequence similarity. The encapsulated genomes are then used to form a weighted bipartite graph, using the weights derived from the similarity matrix. Thus, the problem of matching genes between a pair of genomes can now be transformed to a linear assignment problem, i.e. finding maximum weight matching in a bipartite graph, where the goal is to find the *best* matching for every vertex (protein). Below, we explain in detail the construction of the algorithm.

2.2 Detecting gene families

Let X and Y be two genomes. Here, each genome is essentially a set of protein sequences of the form $X = \{p_1, p_2, p_3, \ldots, p_m\}$ and $Y = \{q_1, q_2, q_3, \ldots, q_n\}$, where the two genomes contain m and n proteins, respectively. The order of the proteins is identical to that identified in their respective chromosomes. We then construct a matrix $M = (M_{ij})$, $1 \leq i \leq m$, $1 \leq i \leq n$ that represents the amino acid sequence similarity between every pair of proteins. In EGM, the similarity between two proteins is based on the alignments returned by BLASTp (Altschul *et al.*, 1990). Two proteins are considered similar only if

the proportion of the alignment coverage on the two protein sequences and its sequence identity is above a user-prescribed threshold t (such as 50%). Where more than one alignment is returned, EGM considers the alignment with the highest identity for scoring purposes. Formally, if l_1 and l_2 are lengths of two protein sequences and k_1 and k_2 are the number of amino acids in their alignment, then $k_1/l_1 \geq t$ and $k_2/l_2 \geq t$. BLASTp is used to generate a list of alignments with a threshold on expect-value ($e < 0.0001$) along with their sequence identity and the corresponding protein sequence start and end points. Note that for any two proteins with no sequence similarity, their corresponding score is initialized to 0. As a result of this step, we build the matrix of similarities M, where any M_{ij} is the similarity between $p_i \in X$ and $q_j \in Y$.

Next, we group together all similar proteins into clusters or pseudo-protein families, adding the notion of families to such data. We term these as *Putative Homology fAMilies* (*PHAMs*). The single-linkage clustering algorithm (Van de Peer, 2004) is used here to form the PHAMs. The measure of similarity acts as a primary filter to eliminate ambiguous edges: further, the associative linkages between proteins increasing the possibility of detecting distant protein family relationships (Catchen *et al.*, 2009; Koonin *et al.*, 1995; Watanabe *et al.*, 1995). Individual PHAMs are identifiable by a label, in our case by an integer. Note here that using stringent thresholds on sequence identity and alignment coverage will result in a high number of clusters (the majority being single member clusters) when comparing distant species. However, in the case of related species, the result will be fewer but more cohesive clusters. Therefore, these thresholds provide a degree of flexibility required for different kinds of analyses.

2.3 Genome encapsulation

A simple encoding method rewrites the genome in an abstract form, where essentially the building blocks for the genomes are transformed from individual proteins to protein families (PHAMs). The resulting encapsulated genomes X' and Y' are now strings of integers, where an integer at any index i is the PHAM label for the protein at the genomic index i. Here, the encapsulated genomes are advantageous as information relating to individual genes and their linkages to other genes across the two species is readily available, while reducing the dimensionality of data. This further helps to visualize gene context and homologous genes across the comparative map. Next, given the encapsulated genomes and the similarity matrix M, we can formulate the problem of finding the best one-to-one matches for each gene between the species as a maximum weight-matching problem in a bipartite graph.

2.4 Maximum weight matching in bipartite graph

Define a weighted bipartite graph, $G = (V \equiv \{X', Y'\}, E)$, where V is the vertex set containing two disjoint sets of nodes representing the PHAM labels in X' and Y', and E is a set of all possible $(m \times n)$ edges between every node $x \in X'$ and $y \in Y'$. Let $W = (w_{ij}), 1 \leq i \leq m, 1 \leq j \leq n$ define matrix of edge weights corresponding to the edges in E. A *matching* in a bipartite graph is a set of edges such that no two edges share a same node (here a node is a protein). A matching, therefore, results in a set of one-to-one matches between the nodes in the bipartition.

2.4.1 Building edge weights In graph matching the similarity matrix M is unsuitable for the direct use as the weight matrix W. This is because unlike alignments (which is equivalent to an order-preserving bipartite matching), the general bipartite graph matching ignores the order of nodes in the two disjoint vertex sets. In an alignment, by the virtue of the strict precedence of nodes, contiguously matched blocks or fragments accumulate a significant score/weight even though the weights on the matched edge are independent of one another. However, this is not the case with a general bipartite graph matching. Therefore, using the similarity matrix M directly as the weight matrix W for E would lead to several spurious node–node correspondences that are possibly non-orthologous and may not form homologous segments between the two genomes. Further, as mentioned earlier, gene strings (or synteny) tend to be conserved, especially is closely related species. As evolutionary distance between two species increases, large-scale gene order shuffling (due to rearrangements) is observed more frequently than in closely related species. Therefore, to prevent spurious correspondences, we construct W in a way that favors matches in a contiguous conserved block of nodes over singleton nodes.

The procedure to compute edge weights in W is as follows: initialize all the edge weights in W to zero. Given the two encapsulated genomes X' and Y', we find all pairs of *maximal substrings* (or fragments), which match exactly between the two encapsulated genomes. We observe that finding sets of contiguous fragments is a common practice in comparative genomics to identify homologous regions (usually adjacent nucleotides), for example, in BLASTZ (Schwartz *et al.*, 2003), PipMaker (Schwartz *et al.*, 2000) and BLAT (Kent, 2002). A hashing technique is applied to efficiently identify such fragments. A hash table is constructed containing positions for all substrings of constant length k (e.g. $k = 3$) for the encapsulated genome X'. To account for inversions, the hash table includes all substrings in the reverse direction of X'. Given the hash index of *k-mers* from X', we now search Y' by sliding across it with a window of size k. This results in a set of matched substrings of length k between the two genomes. We call such substrings *anchors*. We note here that finding fragments using this hashing technique is very efficient: constructing a hash table (for a constant substring length k) takes linear time $O(m)$, where m is the length of the first encapsulated genome X'. Searching against the hash table takes constant time (per search), which is proportional to $O(n)$.

Once the anchors are determined, we extend each identified anchor in both forward and reverse directions to get a set F of *exactly matched maximal substrings*. For each matched maximal fragments in F, of the form $(x_i, \ldots, x_{i+l-1})$ matched with $(y_j, \ldots, y_{j+l-1})$, we compute $\omega = \sum M_{ij} + \ldots + M_{i+l-1, j+l-1}$, where any M_{ij} is the similarity score between the proteins p_i and q_j derived from M. For each matched edge $\{(x_i, y_j), \ldots, (x_{i+l-1}, y_{j+l-1})\}$, the edge weight corresponding to it $\{w_{ij}, \ldots, w_{i+l-1, j+l-1}\}$ is set to ω. Note that to prevent overlapping fragments to be weighted multiple times, edge weight w_{ij} is modified only when ω is greater than the existing value of w_{ij}. Using the above procedure, we derive edge weights in W, which will be used to extract gene correspondences between the two species.

2.4.2 Correspondence extraction Given the bipartite graph G and the weigh matrix W, we extract the set of correspondences using the maximum-weight bipartite graph-matching formulation. Kuhn (1955) initially proposed the Hungarian method for identifying the maximum weight matching in bipartite graphs. We implement the primal-dual linear programming algorithm described by Papadimitriou and Steiglitz (1998), which solves the problem on a graph with $2 \times |V|$ nodes in $O(|V^3|)$ time.

Briefly, the Hungarian method relies on iteratively discovering an *augmenting path* with respect to a current matching C, which is initially empty. A path P of edges $\{u_1, u_2, u_3, \ldots, u_k\}$is considered *augmenting* for a set C if and only if $u_i \in C$ when i is even and $u_i \notin C$ when i is odd, and the sum of weights of all odd-numbered edges in P is greater than the sum of weights of all even-numbered edges in P. When an augmenting path is found, the matching C is replaced by the set of odd-numbered edges in P. The program terminates when no new augmenting path can be found, in which case C gives the maximum weight matching. Figure 2 illustrates an example of the augmentation step. See Papadimitriou and Steiglitz (1998) for full details of the algorithm.

2.5 Evaluation

Evaluation of comparative genomics algorithms is a non-trivial task (Calabrese *et al.*, 2003; Hachiya *et al.*, 2009). Furthermore, direct comparison between tools is difficult due to the variability in output formats. In this article, we will objectively illustrate an estimate of accuracy for the matching produced by EGM, specifically mapping of the Human and Mouse genomes,

which has been a subject to extensive studies (Schwartz *et al.*, 2003). We also demonstrate the results of Human, Mouse, Rat and Zebrafish comparisons using the DAGchainer, ADHoRe and OSfinder tools. Additionally, we calculate the average pairwise similarities between all orthologous pairs, as reported by BLAST. We further conduct a comparison between all four methods against the EnsEMBL compara (Flicek *et al.*, 2008) set of orthologs to estimate precision.

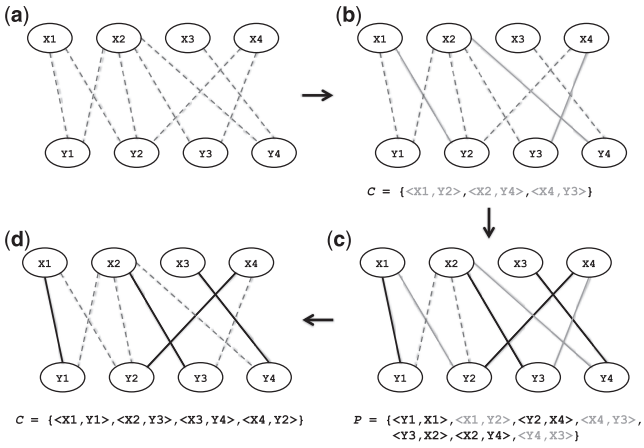

Fig. 2. An idealized illustration of the augmentation step in the Hungarian method. (**a**) Let the dashed edges indicate a set of candidate edges in a bipartite graph from which a maximal bipartite matching (of maximum weight) must be computed. (Hungarian method starts with a complete bipartite graph; assume here that the missing edges are infeasible correspondences. For clarity, the edge weights have been ignored) (**b**) Let the solid grey edges denote a set of current matching C chosen from the candidate edges at some intermediate stage of the algorithm. (**c**) The current set of matchings in set C (shown again solid grey lines) is augmented. P gives the augmenting path for the set of matching in C. The augmenting path alternates between a set of newly chosen edges (shown in black) and edges of C (shown in grey). (**d**) The older set of matchings in C is now replaced with the set of new edges in the alternating path (shown in black). This process (Fig. 2c and d) is iterated until C can no longer be augmented, giving a maximal matching for the bipartite graph.

3 RESULTS

We applied EGM to automatically map homologous gene strings between the Human (Lander *et al.*, 2001), Mouse (Waterston *et al.*, 2002), Rat (Gibbs *et al.*, 2004) and Zebrafish (http://www.sanger.ac.uk/Projects/D_rerio/) genomes. The datasets of protein sequences were obtained from the Integr8 database (Pruess *et al.*, 2005). These sets were sorted according to their encoding gene reference positions (available as gene cross reference data). Proteins without reference positions were omitted from the genome datasets. Details of the dataset and tools are available online from http://vbc.med.monash.edu.au/∼kmahmood/EGM. This results in a total of 21 461 Human, 23 203 Mouse, 22 467 Rat and 22 939 Zebrafish protein-coding genes.

3.1 Detecting orthologs

An application of EGM is to identify the best putative gene orthologs in the process of gene mapping. To evaluate the performance of EGM, we compare the orthologs detected by DAGchainer, ADHoRe and OSfinder. (See Supplementary Methods for these comparisons.)

Table 1 shows the comparisons between the four methods in terms of the number of orthologs identified for each comparison and their average pairwise sequence identity as a measure of accuracy. DAGchainer and ADHoRe often identify matches that belong to more than one variation of the same segment, while EGM calculates the best one-to-one match. Therefore, DAGchainer and ADHoRe results were filtered to include a single match for each gene. See Supplementary Methods for additional details. EGM consistently identifies the highest number of matches in all comparisons, covering a larger proportion of the gene sets. DAGchainer identifies the least number of matching gene pairs (except Mouse/Zebrafish) and the difference was more apparent in the comparisons involving evolutionarily distant Zebrafish genome. Pairwise identity between the matched protein sequences was calculated and matches identified by EGM were found to have the highest identity. In the Human/Mouse, Human/Rat and Mouse/Rat comparisons, all four methods performed well, EGM at 82%

Table 1. Comparison between DAGchainer, ADHoRe, OSfinder and EGM

Comparisons	Ortholog count (average sequence identity)				
	DAGchainer	ADHoRe	OSfinder[a]		EGM
			Total	Similar	
Human–Mouse	11 313 (80%)	13 644 (77%)	15 171	10 781 (70%)	16 214 (79%)
Human–Rat	10 595 (79%)	13 153 (77%)	14 432	9511 (65%)	15 760 (80%)
Human–Zebrafish	4562 (50%)	6532 (53%)	14 077	4971 (40%)	13 649 (59%)
Mouse–Rat	13 242 (84%)	13 481 (86%)	18752	16 329 (68%)	17 799 (88%)
Mouse–Zebrafish	5694 (46%)	6929 (54%)	13 055	4784 (39%)	13 850 (58%)
Rat–Zebrafish	5065 (44%)	6142 (52%)	15 958	6471 (40%)	13 739 (59%)
Average identity	64%	67%		54%	71%

The number of genes matched (ortholog count) is presented for each genome comparison. The average sequence identity for these orthologs is also reported as a measure of accuracy. EGM identifies the highest number of gene matches maintaining a high sequence identity. In terms of ortholog counts, ADHoRe identifies more matches with higher sequence similarity than DAGchainer, especially in the comparisons involving the Zebrafish. OSfinder finds more matches than both DAGchainer and AHDoRe in the Mouse–Rat and Rat–Zebrafish comparisons. Overall, this data suggests that EGM performs the best among the four methods. The advantage of EGM is evident in comparisons involving the more distant Zebrafish genome with the highest count as well as average sequence identity.

[a]An approximate number of gene matches is derived from the OSfinder output. (See Supplementary methods, available online from the EGM website.)

average sequence identity, ADHoRe (80%), DAGchainer (81%) and OSfinder (53%), were able to identify highly similar matches. Expectedly, when comparing more diverged Zebrafish genome in the Human/Zebrafish, Mouse/Zebrafish and Rat/Zebrafish comparisons, the average sequence identity between the matched pairs is lower. Again, as with the gene match count, the difference in match similarity is more significant for the Zebrafish comparisons. EGM matches on average have a 4–15% higher sequence identity than the other tools. Further, on average, EGM identifies orthologs for a significantly larger fraction of the species among all comparisons (see Supplementary File 1).

We have also performed a detailed comparison between all four methods and the EnsEMBL compara orthologs datasets, where the fraction of common orthologs was determined. Considering the overlapping orthologs as true positives, we calculated the *precision* (the fraction of true positives and the total number of reported positives) for each method. For the Human/Mouse, Human/Rat and the Mouse/Rat comparison, DAGchainer (81%), ADHoRe (78%) and EGM (80%) show high precision. However, EGM identifies on average 30% more true positives. The advantage of EGM is more apparent in the comparison of the evolutionarily distant Zebrafish genome. In all of these three comparisons, EGM maintained a significantly higher precision along with the number of true positives. OSfinder performed least favorably in this criterion. (See Supplementary Methods and Supplementary File 2).

Computational efficiency was also assessed. DAGchainer and OSfinder have faster runtime but trades-off for sensitivity. DAGchainer and OSfinder took on average 70–75 min for each comparison. While, EGM took on average 78 min and performs ~10 times faster than ADHoRe (average 860 min). For example, EGM takes ~3 h (complete automated pipeline: pre-processing and correspondence extraction) versus ADHoRe's 19.5 h to perform the Human/Mouse comparison. All comparisons were performed under same conditions on a single 1 GHz processor.

3.2 Application of EGM in detecting conserved gene segments: Human and Mouse comparison

We applied EGM to detect conserved homologous segments and putative orthologs in the Human and Mouse genomes containing 21 461 and 23 203 protein-coding genes. PHAMs were formed using thresholds of 50% on the pairwise sequence identity and alignment overlap. This results in the formation of 13 747 PHAMs. (Summary analysis is available online as Supplementary File 3, available online from the EGM website.) The encapsulated genomes are then formulated, resulting in a string of 21 461 and 23 203 integers for Human and Mouse, respectively. The initial formation of the bipartite graph G formed 391 179 edges, mostly resulting from tandem duplications (Fig. 3 grey dots). Next, gene matching between the two genomes was performed using EGM with seed fragment size set at $k=3$. The output map is represented as a dot plot in Figure 3 (black dots). It is clear from the dot plot that majority of the ambiguous matchings are filtered and only those that reinforce conserved homologous regions across evolutionary rearrangements are mapped (16 214 gene matches).

Further analysis, by combining segments within 3 genes, revealed approximately 845 human gene segments mapped to 1542 homologous segments on the Mouse genome (i.e. collinear strings of conserved genes). Each of the Human and Mouse

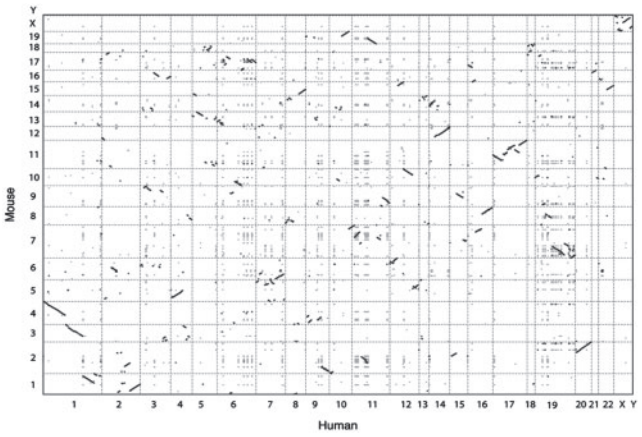

Fig. 3. A dot plot illustration of the Human and Mouse comparison. For each Human gene the corresponding orthologous gene is marked on the Mouse genome. The grids on both axes correspond to the chromosome boundaries. The grey dots represent gene matches as discovered by BLAST hits. Note the number of repeated matchings, where many gene correspondences follow a one-to-many relation (e.g. tandem repeats). The black dots represent the matches determined by EGM. These matches are essentially a subset of the grey dots; however, they represent the best matches for each gene. Further, they clearly depict the conserved syntenic segments between the two species. A higher quality figure is available from the EGM website.

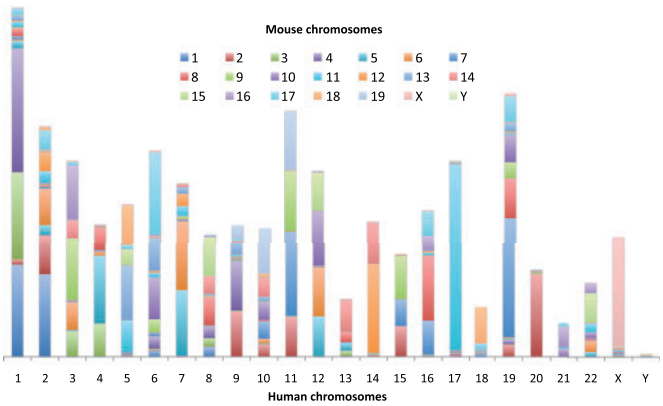

Fig. 4. The bar chart illustrates the distribution of Mouse chromosomes on the Human chromosomes as identified by EGM. In total there were 845 homologous segments from Mouse distributed across a range of Human chromosomes, resulting from complex evolutionary events. These findings agree with studies such as Waterston *et al.* (2002).

gene segments contain on average 11 and 19 genes, respectively. The longest Human segment contains 178 genes, while the longest Mouse segment is 135 genes long. The chromosomal distribution of these segments is illustrated in Figure 4. Furthermore, as many of these segments are only separated by small rearrangements, combining these we obtain 393 gene clusters (Supplementary File 4, available online from the EGM website). We note that Pevzner and Tesler (2003) colleague in their landmark work identified 319 gene clusters forming 281 synteny blocks.

Further analysis was performed by comparing matches produced by the whole genome mapping (genome-wise) against those at the

Human \ Mouse	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	X	Y
1	418/536	26/301	405/501	554/657	31/220	8/194	13/320	36/197	3/300	5/257	23/256	6/120	21/180	3/157	1/130	1/133	29/238	1/76	1/167	1/136	0/6
2	379/436	187/302	1/108	1/135	44/200	164/270	7/171	2/124	6/135	7/140	51/239	89/213	4/97	3/152	2/136	4/70	87/179	12/80	3/95	1/102	0/0
3	1/162	0/155	115/210	3/149	4/146	128/243	5/162	3/114	282/396	0/154	4/185	1/68	0/106	83/178	1/91	247/283	14/148	2/55	0/69	4/133	0/2
4	1/83	1/95	152/198	2/93	312/363	18/104	7/122	111/187	2/112	0/94	1/128	0/47	2/79	2/95	1/61	1/56	2/95	0/70	1/75	1/87	0/3
5	8/101	2/117	1/124	4/123	1/140	0/90	4/132	1/127	0/125	1/118	143/234	2/52	252/284	2/94	74/136	1/46	20/113	189/223	2/64	3/80	0/3
6	33/159	8/178	7/103	48/193	0/100	3/103	9/186	2/91	61/195	187/284	21/170	7/66	162/228	5/105	3/76	9/99	366/476	2/47	1/113	3/99	0/1
7	0/122	4/157	1/99	3/114	296/352	328/399	5/152	3/99	12/131	2/98	46/219	58/159	28/129	11/151	2/101	2/84	1/120	1/49	0/73	1/93	0/3
8	49/106	3/136	37/86	53/140	1/97	1/62	3/110	133/213	3/82	1/81	3/88	0/63	5/59	81/142	187/216	8/42	2/69	0/44	2/57	0/57	0/0
9	4/122	202/271	4/94	224/303	4/80	2/89	4/115	3/87	1/119	4/106	3/109	1/44	63/129	5/66	1/51	0/63	10/105	0/42	60/119	0/58	0/2
10	1/84	59/161	2/83	1/78	1/112	20/106	81/152	7/113	0/82	78/138	5/120	1/46	21/82	83/148	0/76	2/41	0/75	18/75	215/236	1/67	0/0
11	0/168	186/394	3/128	0/114	1/128	0/148	390/621	0/106	281/467	0/206	1/231	0/82	0/90	3/126	4/88	1/130	4/169	1/57	273/338	1/98	0/2
12	1/170	2/182	3/104	1/118	184/258	217/335	4/201	1/114	4/152	260/322	2/237	0/64	0/98	2/91	172/233	4/80	2/127	1/61	1/85	1/79	1/5
13	6/59	0/47	24/77	1/51	36/78	0/37	2/43	44/85	2/54	0/43	0/52	0/26	2/38	145/160	0/29	0/36	1/35	0/42	1/26	5/74	0/0
14	0/109	9/112	1/71	0/81	0/60	0/167	3/125	2/77	1/82	1/81	1/111	404/451	1/87	202/278	1/56	0/61	1/105	0/43	0/89	3/58	0/0
15	0/64	139/194	1/91	1/84	1/65	0/57	120/193	3/72	197/231	0/73	1/80	2/37	0/76	2/44	1/51	0/37	0/67	1/71	0/58	1/47	0/0
16	0/61	3/117	0/60	1/90	6/110	2/85	153/248	301/348	1/102	2/79	9/156	9/45	1/88	0/83	1/79	73/113	115/192	2/56	1/34	2/49	0/3
17	4/138	3/154	0/95	4/120	3/153	3/134	4/166	3/101	0/120	2/127	851/874	3/73	4/89	1/83	1/145	0/88	3/116	0/59	1/64	3/93	0/1
18	19/53	0/67	1/30	0/38	2/42	0/32	0/45	1/56	1/41	0/39	2/47	1/36	5/46	3/35	0/32	0/20	28/56	165/168	0/20	0/23	0/0
19	4/162	58/202	2/139	19/182	5/141	4/154	558/693	184/275	78/241	125/235	4/175	8/95	40/153	2/129	2/98	1/92	116/257	2/63	2/121	3/97	1/3
20	1/63	379/389	6/55	0/73	3/59	1/53	1/66	0/79	2/59	2/62	0/63	2/60	1/47	0/51	0/63	0/25	1/52	1/50	1/40	0/43	0/2
21	0/17	0/11	0/11	0/11	0/18	0/10	0/17	0/15	0/35	36/48	3/22	0/8	0/8	0/9	0/10	95/102	17/34	0/4	0/6	0/8	0/0
22	1/36	0/43	0/36	3/34	21/78	51/93	1/62	6/39	2/49	29/64	37/100	2/63	1/36	2/72	125/145	44/76	2/43	1/24	0/24	2/37	0/1
X	5/108	2/91	3/98	2/99	6/96	1/103	4/116	3/111	1/119	5/98	4/122	2/50	2/83	5/88	4/59	1/59	2/81	2/57	1/49	486/500	0/7
Y	2/6	0/7	0/5	0/6	0/7	1/7	0/7	1/6	0/4	3/7	0/6	0/4	2/4	0/3	0/3	0/2	2/5	1/5	0/0	4/14	5/6

<9	<26	<50	<100	<150	<200	<250	<300	<350	<390	>=390

Fig. 5. This Oxford grid shows gene orthologs in syntenic regions among Human and Mouse as detected by EGM. The putative orthologs are represented as 'a/b' and are described as the ortholog count attained from (**a**) when comparison is performed between the Human/Mouse as a whole and (**b**) from the chromosome-by-chromosome level matchings. By chromosome-by-chromosome matching, we mean that $n_1 \times n_2$ different EGM runs were performed for each pair of chromosomes in two genomes (containing n_1 and n_2 chromosomes, respectively). The significant overlap of the matching genes in the conserved segments suggests that EGM is both sensitive and accurate for large-scale comparisons. Tools are also available from the EGM website that can be used to derive such data.

individual chromosome level (chromosome-wise). Chromosome-wise maps were produced by running EGM between each pair of Human and Mouse chromosomes. (see Supplementary File 5, available online from the EGM website.) Next, the whole Human/Mouse genome-wise gene matches were split at the chromosome level. The overlap of the two sets of matchings (genome- and chromosome-wise) is calculated and the results are summarized as an Oxford grid (Edwards, 1991) in Figure 5. The overlap indicates, for example, at the genome level 486 orthologs are detected between the X chromosomes of the two species, while the mapping of the individual X chromosomes reveals 500 orthologs. Similarly, 851 orthologs were detected at the genome-wise comparison between Human chromosome 17 and Mouse chromosome 11, again showing a significant overlap with 874 orthologs detected in the chromosome-wise comparison. (See intersection dot plot in Fig. 6.) Conversely, the comparison between the Human chromosome 1 and Mouse chromosome 2 results in 301 matches. However, in the genome-wise comparison, there are only 26 overlapping genes, with the rest finding better matches. Supplementary files 6 and 7 provide Oxford grids and dot plots for the Human versus Rat and Human versus Zebrafish comparisons, respectively, produced from the EGM results.

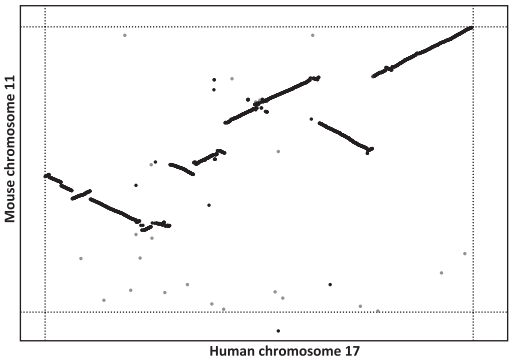

Fig. 6. Human chromosome 17 versus Mouse chromosome 11. Dot plots representing the accuracy of EGM in mapping large-scale genomes with complex evolutionary rearrangement events. The dot plot represents the combination of gene matches identified by EGM comparisons at a chromosome level (red dots) and at a genome level (grey dots, essentially magnified from Fig. 3). The intersection of the two datasets is shown in dark dots. This plot clearly shows that EGM is sensitive and robust in identifying complex genome rearrangements and accurate gene matching at various data scales.

3.3 Implementation

EGM is implemented as a fully automated, standalone application and works similar to a simple molecular alignment program. It takes as input two genomes (FASTA formatted protein sequences) and produces a gene-by-gene matching along with a dot plot. A series of helpful tools are also provided that assist in setting up the inputs as well as the analysis of the results. We also make available the complete chromosome-wise map for the Human and Mouse genomes produced by EGM (available online). Users can choose the chromosomes of interest to analyze the maps.

4 DISCUSSION

EGM presents a further step towards fully automating the complex process of gene-by-gene matching. EGM is a powerful tool capable of identifying the best one-to-one gene correspondences between a pair of genomes that leads to the identification of conserved homologous segments, a crucial step in the identification of syntenic regions. We have shown the EGM tool to be a useful method that integrates several pieces of information through the pipeline to produce a global comparative map. A series of complex genome comparison, including a Human and Mouse comparison, were

performed using EGM. The results clearly show its effectiveness in providing a detailed comparison. The results include information about the putative gene orthologs, protein families, their context and organization on the genomes. Our experiments show that EGM is able to sensitively identify conserved gene segments and reveal complex evolutionary rearrangement events in large-scale (species-level) and small-scale (chromosome-level) data. The combination of data provided by EGM comparison is a useful tool in the prediction of protein function. With respect to conserved gene segments, encapsulated genomes allow easier detection of identical cluster labels rather than individual proteins. Encapsulation also makes it easy to comprehend gene context and topology information, especially in highly conserved regions; we believe this has a potential to help researchers explain their functional significance and possible interactions.

Comparison with DAGchainer, ADHoRe and OSfinder revealed EGM's ability to consistently detect more orthologs while maintaining high similarity. DAGchainer works by utilizing the diagonal properties from the anchors, and similar to the findings of Hachiya *et al.* (2009), the precision to identify accurate orthologs diminishes when comparing distant species (such as the comparisons involving the Zebrafish) as non-orthologous anchors significantly outnumber the orthologous anchors in the input; therefore, resulting in fewer gene matches in smaller syntenic regions. ADHoRe, a quality-based method, determines orthologous segments by iteratively increasing the inter-segmental distance threshold for sets of anchors that fit diagonal linear regressions. The quality of these diagonals is assessed by the goodness of their fit. Overall, DAGchainer, ADHoRe and OSfinder performed well in the comparisons between the Human, Mouse and Rat genomes, but to a lesser degree in comparisons involving more distantly related organisms such as the Zebrafish genome. However, from a point of view of the methodology, a practical limitation with DAGchainer and ADHoRe is their reliance on quality parameter scoring schemes that are not well characterized and difficult to ascertain when little is known about the genomes being compared. Further, DAGchainer and ADHoRe gene matches reveal many-to-many relationships, and in many cases this leads to many variations of the same conserved segments being reported (Soderlund *et al.*, 2006). The EnsEMBL compara database tool curates multi-species comparative data. We utilized these Compara-curated orthologs to evaluate the precision of the four methods mentioned here. EGM showed high precision in detecting true positive orthologs for species at varying evolutionary distances.

Identification of best one-to-one gene matches between a pair of genomes is a pivotal step in characterizing unknown proteins, as such relationships help identify genes with common ancestors (Sankoff, 1999). In the case of EGM, the identification of best orthologs relies on sequence similarity and conservation of genome context and not on any quality-based measure or threshold. The use of similarity scores in detecting functional orthologs remains a preferred method used in several studies, and was found to perform significantly well in comparison to other more sophisticated techniques (Altenhoff *et al.*, 2009). Further, as mentioned earlier, one-to-one orthologs are important in several studies ranging from protein–protein interactions in multiple species to identification of biological pathways. However, as thousands of new proteins are sequenced from increasingly complex genomes, the task of assigning orthologs is becoming more complex. In addition, the presence of large protein families (paralogs), evolutionary processes such as speciation and gene duplications or both are not uncommon (Bandyopadhyay *et al.*, 2006; Dehal *et al.*, 2005; Sjolander, 2004; Wu *et al.*, 2006). Therefore, it is a non-trivial, yet important task, to identify co-orthologs, as opposed to only the 'best' orthologs, and homologous segments to paint a more comprehensive evolutionary picture. As future directions, we are pursuing to evolve our EGM approach to perform this task. Deploying an iterative strategy, where at every iteration, a new set of matching can be reported, while eliminating the possibility of identifying the previously reported matching. We believe, collectively from all iterations we can extract sets of co-orthologs and syntenic regions.

A bottleneck that remains in most of the current methods is the reliance on external tools (such as BLAST) to assign initial similarity relationships. This is often time consuming and involves manual processing. We believe that alignment-free methods (Vinga *et al.*, 2003) can be employed to significantly accelerate this task with reasonable sensitivity. We are also pursuing this as a future direction. Similarly, visualizing comparative maps remains complicated. Much of these data are generated as long text files unrolling complex information. Recently, resources such as the Synteny Database (Catchen *et al.*, 2009) along with other web technologies such as HTML5, have demonstrated their usefulness to simplify the visualization. Development of efficient visualization tools can greatly improve the analysis of such data.

5 CONCLUSION

Gene similarity, gene context and order are a common theme when searching for functionally related genes and overall conservation of proteins. Analyzing these properties often requires manual intervention (Altenhoff *et al.*, 2009; Barrangou *et al.*, 2009; Pereyre *et al.*, 2009). We believe that EGM with its modular approach is able to help automate these tasks. We have found EGM to be a powerful method that identifies gene matching at the one-to-one level. The EGM tool is a simple to use and fully automated. Our results indicate that EGM has a great potential in its competitive performance in comparison with other popular approaches based on comparisons between several species. The resulting comparative maps have also shown EGM to be sensitive in detecting complex rearrangement events. In addition, the output produced by EGM clearly depicts how individual genes match across a pair of genomes, their genomic context and protein family information. EGM is able to perform two major integrated tasks: (i) finding the best ortholog for each gene and (ii) identifying homologous segments shared between the compared species.

ACKNOWLEDGEMENTS

J.C.W is an ARC Federation Fellow and Honorary NHMRC Principal Research Fellow. A.M.B is an NHMRC Senior Research Fellow. J.S is an NHMRC Peter Doherty Fellow. K.M is an ARC PhD student. We thank Noel Faux and Ruby Law for their discussions. We thank the Monash e-Research Centre and the Victorian Bioinformatics Consortium for computational resources.

Funding: Grants from the Australian Research Council and the National Health and Medical Research Council of Australia; National ICT Australia (NICTA) is funded by: the Australian

Government's Department of Communications, Information Technology and the Arts; the Australian Research Council through Backing Australia's Ability and ICT Centre of Excellence program.

Conflict of Interest: none declared.

REFERENCES

Abouelhoda,M.M.M.I. (2005) Algorithms and a software system for comparative genome analysis. PhD Thesis, Theoretical Computer Science Department. University of Ulm, Ulm, pp. 191.

Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ayala,J.A. *et al.* (1994) New comprehensive biochemistry. In *Bacterial Cell Wall*. Elsevier Science, London.

Bandyopadhyay,S. *et al.* (2006) Systematic identification of functional orthologs based on protein network comparison. *Genome Res.*, **16**, 428–435.

Bansal,A.K. *et al.* (1998) Automated pair-wise comparisons of microbial genomes. *Math. Model. Sci. Comput.*, **19**, 1–23.

Barrangou,R. *et al.* (2009) Comparison of the complete genome sequences of Bifidobacterium animalis subsp. lactis DSM 10140 and Bl-04. *J. Bacteriol.*, **191**, 4144–4151.

Blanchette,M. (2007) Computation and analysis of genomic multi-sequence alignments. *Annu. Rev. Genomics Hum. Genet.*, **8**, 193–213.

Blanchette,M. *et al.* (2009) Genome-wide analysis of alternative pre-mRNA splicing and RNA-binding specificities of the Drosophila hnRNP A/B family members. *Mol. Cell*, **33**, 438–449.

Calabrese,P.P. *et al.* (2003) Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, **19** (Suppl. 1), i74–i80.

Catchen,J.M. *et al.* (2009) Automated identification of conserved synteny after whole-genome duplication. *Genome Res.*, **19**, 1497–1505.

Chain,P. *et al.* (2003) An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Brief. Bioinform.*, **4**, 105–123.

Dandekar,T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.

Das,S. *et al.* (2009) ABWGAT: anchor-based whole genome analysis tool. *Bioinformatics*, **25**, 3319–3320.

Dehal,P. and Boore,J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, **3**, e314.

Dewey,C.N. *et al.* (2006) Parametric alignment of Drosophila genomes. *PLoS Comput. Biol.*, **2**, e73.

Edwards,J.H. (1991) The Oxford Grid. *Ann. Hum. Genet.*, **55**, 17–31.

Flicek,P. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.

Fu,Z. *et al.* (2007) MSOAR: a high-throughput ortholog assignment system based on genome rearrangement. *J. Comput. Biol.*, **14**, 1160–1175.

Gibbs,R.A. *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Goldberg,D.S. *et al.* (2000) Algorithms for constructing comparative maps. In Sankoff,D. and Nadeau,J.H. (eds) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families*. Kluwer Academic Press, Dordrecht, The Netherlands, pp. 243.

Haas,B.J. *et al.* (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, **20**, 3643–3646.

Hachiya,T. *et al.* (2009) Accurate identification of orthologous segments among multiple genomes. *Bioinformatics*, **25**, 853–860.

Himmelreich,R. *et al.* (1997) Comparative analysis of the genomes of the bacteria Mycoplasma pneumoniae and Mycoplasma genitalium. *Nucleic Acids Res.*, **25**, 701–712.

Huynen,M. *et al.* (2000a) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.

Huynen,M. *et al.* (2000b) Exploitation of gene context, *Curr. Opin. Struct. Biol.*, **10**, 366-370.

Kalafus,K.J. *et al.* (2004) Pash: efficient genome-scale sequence anchoring by Positional Hashing. *Genome Res.*, **14**, 672–678.

Kellis,M. *et al.* (2004) Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J. Comput. Biol.*, **11**, 319–355.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Kent,W.J. *et al.* (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.

Kent,W.J. and Zahler,A.M. (2000) Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. *Genome Res.*, **10**, 1115–1125.

Koonin,E.V. *et al.* (1995) Sequence similarity analysis of Escherichia coli proteins: functional and evolutionary implications. *Proc. Natl Acad. Sci. USA*, **92**, 11921–11925.

Kuhn,H.W. (1955) The Hungarian Method for the assignment problem. *Nav. Res. Logistics Q.*, **2**, 83–97.

Kurtz,S. *et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.

Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lathe,W.C., 3rd *et al.* (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci.*, **25**, 474–479.

Lemoine,F. *et al.* (2007) Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC Evol. Biol.*, **7**, 237.

Nadeau,J.H. and Taylor,B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA*, **81**, 814–818.

Papadimitriou,C.H. and Steiglitz,K. (1998) *Combinatorial optimization: algorithms and complexity*. Courier Dover Publications, Mineola, NY.

Peng,Q. *et al.* (2006) The fragile breakage versus random breakage models of chromosome evolution. *PLoS Comput. Biol.*, **2**, e14.

Peng,Q. *et al.* (2009) Decoding synteny blocks and large-scale duplications in mammalian and plant genomes. In *Algorithms in Bioinformatics*. Springer, Berlin/Heidelberg, pp. 220–232.

Pereyre,S. *et al.* (2009) Life on arginine for Mycoplasma hominis: clues from its minimal genome and comparison with other human urogenital mycoplasmas. *PLoS Genet.*, **5**, e1000677.

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

Pruess,M. *et al.* (2005) The Integr8 project—a resource for genomic and proteomic data. *In Silico Biol.*, **5**, 179–185.

Rasmussen,M.D. and Kellis,M. (2007) Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes, *Genome Res.*, **17**, 1932–1942.

Rogozin,I.B. *et al.* (2004) Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. *Brief. Bioinform.*, **5**, 131–149.

Salse,J. *et al.* (2009) Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief. Bioinform.*, **10**, 619–630.

Sankoff, D. (1999) Genome rearrangement with gene families, *Bioinformatics*, **15**, 909–917.

Sankoff, D. and Nadeau, J.H. (2000) Comparative Genomics: Emperical and analytical approaches to gene order dynamics, map alignment and evolution of gene families. In Dress, A. (ed), *Computational Biology Series*. Kluwer Academic Publishers.

Schwartz,S. *et al.* (2000) PipMaker—a web server for aligning two genomic DNA sequences, *Genome Res.*, **10**, 577–586.

Schwartz,S. *et al.* (2003) Human-mouse alignments with BLASTZ, *Genome Res.*, **13**, 103–107.

Sharan,R. *et al.* (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.

Shi,G. *et al.* (2010) MSOAR 2.0: incorporating tandem duplications into ortholog assignment based on genome rearrangement. *BMC Bioinform.*, **11**, 10.

Sjolander,K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170–179.

Soderlund,C. *et al.* (2006) SyMAP: a system for discovering and viewing syntenic regions of FPC maps. *Genome Res.*, **16**, 1159–1168.

Swidan,F. *et al.* (2006) An integrative method for accurate comparative genome mapping. *PLoS Comput. Biol.*, **2**, e75.

Szklarczyk,R. and Heringa,J. (2006) AuberGene–a sensitive genome alignment tool, *Bioinformatics*, **22**, 1431-1436.

Tamames,J. (2001) Evolution of gene order conservation in prokaryotes. *Genome Biol.*, **2**, 1–11.

Tamames,J. *et al.* (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, **44**, 66–73.

Van de Peer,Y. (2004) Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.*, **5**, 752–763.

Vandepoele,K. *et al.* (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Res.*, **12**, 1792–1801.

Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison-a review. *Bioinformatics*, **19**, 513–523.

Wang,Z. *et al.* (2009) CHSMiner: a GUI tool to identify chromosomal homologous segments. *Algorithms Mol. Biol.*, **4**, 2.

Watanabe,H. and Otsuka,J. (1995) A comprehensive representation of extensive similarity linkage between large numbers of proteins. *Comput. Appl. Biosci.*, **11**, 159–166.

Waterston,R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Wu,F. *et al.* (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics*, **174**, 1407–1420.