Recent developments in computational approaches for uncovering genomic homology

Cedric Simillion, Klaas Vandepoele, and Yves Van de Peer*

Summary

Identifying genomic homology within and between genomes is essential when studying genome evolution. In the past years, different computational techniques have been developed to detect homology even when the actual similarity between homologous segments is low. Depending on the strategy used, these methods search for pairs of chromosomal segments between which either both gene content and order are conserved or gene content only. However, due to fact that, after their divergence, homologous segments can lose a different set of genes, these methods still often fail to detect genomic homology. Recently, more advanced approaches have been developed that can combine gene order and content information of multiple genomic segments. ***BioEssays*** **26:1225–1235, 2004.**

Introduction

Ever since man started to study the variety of living organisms that surrounded him, it was noted that many different species share structures that, while at first sight they might look vastly different, are remarkably similar in anatomical detail. For instance, the arms of a human share the same skeletal components as the wings of a bat. However, it was not until Charles Darwin that it was realized that this similarity is caused by the fact that organisms are descendants from the same, nowadays extinct, ancestral species. This similarity through common ancestry has been referred to as homology and it was soon recognized that detecting these homology relationships is essential when studying the evolutionary history of organisms. It was also noted that homologies could even exist within a single organism, as for example the front, middle and hind legs of a grasshopper. As the science of biology advanced, more sophisticated techniques allowed uncovering homologies that were previously unperceivable. For example, the advent of microscopy allowed the identification of similar tissue and cell types, such as nerve or muscle cells in almost all animals. Or, at an even higher level of detail, electron microscopy revealed that all eukaryote species share a similar, double-membraned, subcellular structure.

Over the last few decades, molecular techniques have been developed that allow biologists to study and compare organisms at the highest level of resolution. Today, the availability of an increasing amount of genome sequence information from a large variety of organisms makes possible the study of homology at the gene and genome level. Just as with anatomical structures, genomic homology can be observed between different organisms as well as within the same organism. Indeed, related species often share large genomic segments[1–3] that are similar while, at the same time, remnants of block duplication events or even entire genome duplications might be identified within the same genome. Needless to say, the identification of homologous genomic regions is thus an essential prerequisite when studying the evolution of genomes, both within and between organisms. Identifying intergenomic homology thus allows researchers to assess the impact of rearrangement events,[4–8] while intragenomic homology provides insight into the duplication history of a genome.[9–11]

Due to different types of rearrangements, gene duplication and gene loss, the identification of genomic homology is not always obvious. However, here too, more advanced techniques have recently allowed the identification of previously undetectable homologies. In this review, we will discuss the different computational techniques that are currently available for the detection of genomic homology, even when the actual similarity between genomic segments is low or at first sight nonexistent.

The map-based approach

Intuitively, the most straightforward way to uncover homologous genomic segments would be to perform an all-against-all comparison of chromosomal DNA sequences. Although recently several computational tools have become available for the sequence alignment of entire chromosomes, or large parts thereof,[12–18] this approach will only reveal recently diverged segments. Indeed, as mutations accumulate during evolution, homologous segments become hard to identify as such when only considering the primary DNA sequence.

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Belgium
*Correspondence to: Yves Van de Peer, Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium.
E-mail: yves.vandepeer@psb.ugent.be
DOI 10.1002/bies.20127
Published online in Wiley InterScience (www.interscience.wiley.com).

Consequently, this strategy is less suited when one wishes to search for ancient homology relationships. Instead, a better way to detect similarity between two genomic segments is by comparing their overall gene content and, optionally, gene order.[19] Two chromosomal segments can then be considered homologous if they share a significantly higher number of homologous genes than what would be expected by comparing non-related genomic regions. Of course, this means that a reasonably good structural gene annotation must be available, although this is rarely a problem, since most, if not all, genome sequencing efforts also provide an annotation. Historically, this strategy was used in mapping studies long before computational methods and complete genome sequences were available.[1,3,20] For this reason, the detection of chromosomal homology by comparing gene content and order between different segments is commonly referred to as the map-based approach.

There are different ways to implement such a map-comparing strategy, depending on whether one wants to take into account both gene content and order or gene content only. However, usually, the basic concept of these different implementations is to consider a chromosome as a list of genes, sorted in the order in which they appear on that chromosome. Starting from these lists, a first crucial step in the implementation of a map-based approach is the identification of homologous genes between chromosomes. Usually this is done by performing an all-against-all similarity search with the protein sequences of those genes (e.g. using BLAST[21]).

When considering conservation of both gene content and order, the most-common approach is to represent the results of the similarity search in a dot matrix. In such a matrix, which we shall further refer to as a gene homology matrix or GHM, the rows and columns correspond to the positions of the genes along the respective chromosomes. A cell will now contain a non-zero value if the genes of the corresponding row and column turn out to be homologous. Optionally, such a cell can be marked as positive when both genes are transcribed on the same DNA strand or as negative if they have opposite transcriptional orientations. As a result of this matrix representation, pairs of duplicated segments become visible as a series of diagonally arranged non-zero elements (referred to as 'dots') in the matrix. Fig. 1 gives a hypothetical example of a GHM. Consequently, homologous regions can be delineated computationally by grouping such series of diagonally arranged dots.

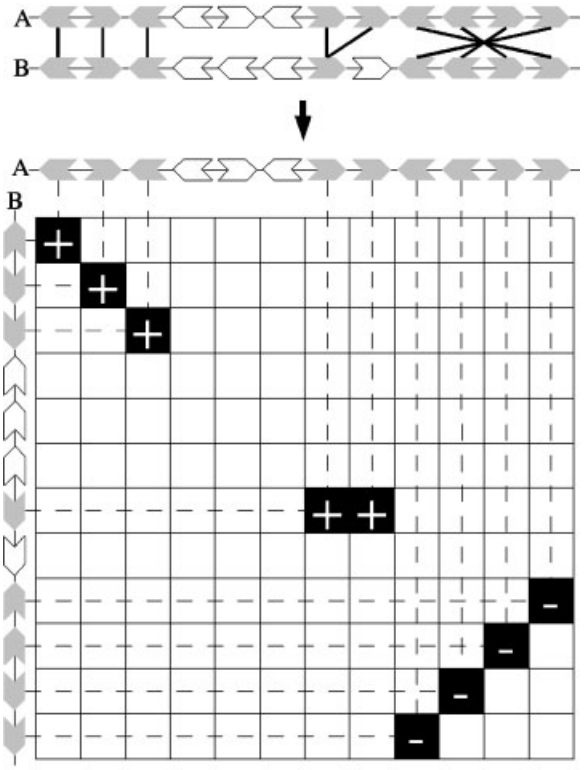

Figure 1. Hypothetical gene homology matrix. The top section shows two colinear genomic segments A and B. Every arrowhead denotes a gene with the orientation representing the transcriptional orientation of the genes. Genes shown in grey and connected by lines represent pairs of homologous genes. The bottom section shows the actual GHM derived from these two segments. Every column represents a gene of segment A while every row corresponds to a gene of segment B. A cell will contain a non-zero value (i.e. is marked in black) if the gene products of the corresponding row and column are homologous to each other. In addition, a '+' indicates that both genes have the same transcriptional orientation while a '–' signifies opposite orientations. Note that pairs of homologous segments become apparent as series of diagonal elements, with the direction depending on the relative orientation of both segments.

First applications of the map-based approach

One of the very first studies that was based on the construction and analysis of GHMs appeared in 1997 when Wolfe and Shields[22] analyzed the complete yeast genome (*Saccharomyces cerevisiae*) to find duplicated segments. Here, the authors took into account transcriptional orientation of the duplicated gene pairs and imposed rather strict criteria to define homologous regions. With this strategy, it was concluded that at least 50% of the yeast genome was covered by duplicated regions. Using a statistical test, they showed that the number of duplicated regions in yeast is significantly higher than what would be expected by chance if the duplicated gene pairs would have arisen individually. Almost no triplicated regions were observed. Furthermore, the orientation towards the centromere of all the detected duplicated segments was generally conserved. These two observations led the authors to conclude that the detected duplication pattern could only be obtained by a single large-scale duplication event, most likely a tetraploidisation event. Phylogenetic analysis furthermore

showed that this event must have occurred after the divergence of the *Saccharomyces* en *Kluveromyces* lineages.[22].

Three years later Vision and coworkers[23] applied a similar approach to the genome of the model plant *Arabidopsis thaliana*. Even before the completion of its sequence, several studies already indicated that, despite of its small size, a large portion of this genome was duplicated.[24,25] When the sequence was completed, the *Arabidopsis* Genome Initiative (AGI) reported that 60% of the genome resided in duplicated regions but that no multiple (i.e. more than two) duplications had occurred.[26] Consequently, the AGI concluded that the observed duplication pattern could only be explained by a single duplication event, which was, given the total size of the duplicated fraction, probably also a tetraploidisation event. However, this analysis was done using direct comparison of the primary DNA sequence. As mentioned before, this kind of comparison can only reveal relatively recently duplicated regions.

Vision and coworkers used the map-based approach to develop an algorithm that searches for duplicated regions in the *Arabidopsis* genome. Their strategy starts again from a GHM on which the transcriptional orientation of the gene pairs is indicated. When constructing this GHM, series of tandemly duplicated genes were remapped onto the row or column with the lowest index, i.e., the first gene of the tandem array, in order to reduce the distortion of GHM diagonals by horizontal or vertical patterns produced by many tandem duplications (see also Fig. 1). By considering the dots of the obtained GHM now as nodes in a graph and by assigning weights to the vertices connecting them, diagonal series of dots were detected as minimum-weight paths. These series were then combined to delineate pairs of homologous, in this case duplicated, regions. Using this method, Vision and coworkers concluded that as much as 81% of the *Arabidopsis* genome was duplicated and that numerous regions had also undergone multiple duplications.[23] These observations clearly suggested that the genome of *Arabidopsis* was shaped by multiple rounds of large-scale gene duplication.

This increase in sensitivity clearly illustrates the superiority of the map-based approach compared to direct DNA comparisons for the detection of homologous regions. However, the method of Vision and coworkers[23] has also several drawbacks. The minimum-weight path-finding routine can generate overlapping paths such that the results require careful manual inspection afterwards. Also, the weight-function used is governed by three user-defined parameters that have no real biological meaning such that choosing the proper parameter values is arbitrary.

General implementations of the map-based approach

While each of the former two methods[22,23] was especially developed to study specific genomes (*Saccharomyces* and *Arabidopsis* respectively), later implementations provided more generally applicable approaches and also overcame some of the drawbacks of the first methods. In 2002, we published a software tool called ADHoRe that automatically detects pairs of homologous regions and performs statistical evaluation of these pairs.[27] This method also starts by constructing a GHM where the transcriptional orientation of the genes involved is considered and where tandem genes are remapped. Again, pairs of homologous segments appear as series of diagonal dots. However, ADHoRe uses a special distance function that yields a shorter distance for elements that are in diagonal close proximity than elements that are in horizontal or vertical proximity.[27] Furthermore, the algorithm is able to deal with both small-scale and large-scale inversions. The entire procedure is controled by only two parameters: a gap size denoting the maximum allowed distance between the dots in a cluster and a minimum quality factor that corresponds to the minimum degree of conservation of gene order. For each cluster, the statistical significance is evaluated individually by comparing both the total number of dots (i.e. homologous gene pairs) and their average distance with clusters derived from randomized datasets. The sensitivitiy of this method was illustrated by a comparative analysis of the *Arabidopsis* and rice (*Oryza sativa*) genomes.[27] Although the rice genome was still very incomplete at that time, ADHoRe was able to detect several colinear stretches between these two genomes, showing that, using a map-based approach, one is able to uncover genomic homology even across the monocot–dicot split.

Recently, Calabrese and coworkers published the FISH (Fast Identification of Segmental Homologies) software package,[28] which is an improved implementation of the method used by Vision and coworkers in 2000[23] and is similar to the ADHoRe software.

All methods mentioned above define homologous regions by detecting diagonal patterns in a GHM. A notable exception to this is the LineUp package from Hampson and coauthors.[29] This algorithm detects homologous segments as runs of colinear markers or genes by starting from pairs of homologous genes between two chromosomes and extending the runs by looking for subsequent pairs of homologous genes. As with the other methods, this algorithm also allows for gaps between such subsequent pairs and small rearrangements of the gene order. Additionally, the method is also able to deal with unresolved distances between two neighboring markers on the same segment. This is usually not an issue when dealing with physical, sequence-derived maps but is of considerable relevance when dealing with genetic maps that typically have a limited mapping resolution. The statistical significance of the individual detected runs is evaluated using randomization tests and by considering the number of homologous gene pairs in a run and the length of the run. When applying their method to different genetic maps of the maize

(*Zea mays*) genome, the authors found that a considerable fraction of the maize genome is duplicated thereby reconfirming previous studies that indicate that maize is an ancient allotetraploid.[30]

Discarding gene order

While methods that look for conserved gene content and order have proven to be sensitive enough to recover traces of ancient homology or duplication events, sometimes the requirement for conserved gene order turns out to be too strict. Indeed, when analyzing the human genome with a tool like ADHoRe, little evidence for ancient large-scale gene duplication is found (unpublished results). However, by releasing the constraint of conserved gene order, Abi-Rached and coworkers[31] have found considerable evidence for duplicated segments in the human genome associated with the MHC loci. Likewise, Larhammar and coworkers[32] found that the regions flanking the HOX loci of different vertebrate genomes were conserved in gene content but not in order, thereby providing additional evidence that the HOX loci arose through a large-scale gene duplication event. These studies relied on manually identifying the duplicated regions. Several authors however have already devised automated ways to look for conserved gene content.

A first way of evaluating conserved gene content (but not order) in an automated manner is to compare two genomic windows and to count the number of homologous gene pairs between these windows. This strategy has been adopted by Hughes and Friedman[33] when they analyzed the genomes of *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*. In this analysis, the authors defined a genomic window as a region containing eight non-single-copy genes (i.e. genes with at least one homolog somewhere in the genome). All possible, non-overlapping windows are then compared with each other. By comparing the results from real genomes with those from randomized genomes, it was shown that all three genomes investigated showed a significantly larger number of windows sharing at least two homologous gene pairs more than would be expected under random gene distribution. These results also indicated that the detected duplication pattern could vary greatly between different genomes. For example, the genome of *C. elegans* contained only intrachromosomal segmental duplications whereas, in *S. cerevisiae*, the vast majority of detected duplications was interchromosomal. Additionally, by calculating the fraction of synonymous substitutions between the pairs of duplicated genes, it was shown that, apart from the ancient polyploidy event described by Wolfe and Shields,[22] other duplication events might also have occurred in the yeast genome (see also Koszul and coworkers[34]). A limitation of the method used, however, is that it only detects the global presence of duplicated segments in a genome but does not delineate individual pairs of homologous segments.

A modified version of this algorithm was developed by Cavalcanti and coworkers.[35] Contrary to Hughes and Friedman, these authors considered all possible genomic windows in a genome rather than dividing the dataset into non-overlapping windows. Next to this, the requirement was made that, for every pair of matching windows, all homologous genes in one window should have a counterpart in the other window. Additionally, they also provided the option of also taking into account conservation of gene order, with or without preservation of transcriptional orientation. With these adaptations, the authors found fewer duplicated segments in *Saccharomyces* but more in *Caenorhabditis*, although the general patterns of duplication observed were similar to those of Hughes and Friedman.

In a large-scale analysis of the human genome, McLysaght and coworkers[36] used a different approach that was developed by Hokamp.[37] Contrary to the method discussed above, this algorithm is able to identify and delineate individual pairs of homologous segments. The input is a complete list of similarity matches between all genes in the dataset. Starting from two homologous genes, each on a different chromosomal location, the software looks for two other homologous genes that are each located within a pre-defined distance of the former two genes. If such a pair is found, it is added to the first pair to form a cluster of homologous genes. Next, additional pairs of genes are searched for that are in the vicinity of the genes already in the cluster and subsequently added. This is iterated until no more additional pairs can be added to the cluster. The resulting clusters are then used to delineate pairs of homologous segments or ''paralogons''.[38] Using this strategy, McLysaght and coworkers found that 44% of the human genome was covered by paralogons with six or more pairs of duplicated genes. Combined with phylogenetic dating, this observation prompted the authors to conclude that a polyploidy event is likely to have happened early in the origin of vertebrates.

Recently, Hampson and coworkers published CloseUp, a software tool that detects conservation of gene content.[39] The algorithm is similar to the method of Hokamp but only extends a detected cluster if the number of homologous gene pairs is significantly greater than would be expected by chance. Also, the statistical validation of the detected clusters takes into account both the number of homologous gene pairs and the distance between these genes on the respective chromosomes whereas the method of Hokamp[37] considers the number of gene pairs.

Extending the map-based approach

The drawback of considering only conservation of gene content is that one needs more pairs of homologous genes between two segments to obtain the same statistical significance as when gene order is conserved as well. Durand and Sankoff[40] show that detecting a stretch of *m* genes in the

same order is *m!* times as significant as finding these *m* genes in any order. Moreover, the fact that a pair of anciently duplicated segments cannot be recognized as such is not necessarily caused by changes of gene order in the segments. It has indeed been observed that, after duplication, two daughter segments lose a different, complementary set of genes.

This phenomenon, referred to as 'differential gene loss' was first described by Ku and coworkers.[41] By manually comparing the gene order of a tomato BAC with the genomic scaffolds of *Arabidopsis*, four different parts of the latter genome were found to be colinear with the BAC. Interestingly, it was observed that the homologous counterparts of the tomato (*Solanum lycopersicum*) BAC genes were scattered throughout the four *Arabidopsis* segments but that the order in which these genes occurred on these segments was still conserved compared to the BAC segment.

A similar pattern was observed at an even more dramatic scale when comparing the *Arabidopsis* genome with that of rice.[42] We noticed that two different *Arabidopsis* segments showed significant colinearity with the same region in rice from which it could be concluded that both segments were homologous. However, colinearity between these two *Arabidopsis* segments could not be observed, the reason being that the two segments did not share a single homologous gene pair anymore due to extensive differential gene loss. Such a pair of duplicated segments for which homology can only be inferred through comparison with another genome is called a 'ghost duplication'. Later, we were able to identify additional examples of ghost duplications, indicating that differential gene loss is a commonly occurring phenomenon. Moreover, we can also detect such a pattern of differential gene loss by comparing *Arabidopsis* segments only. In this case, we refer to these duplications as 'hidden' duplications[43] indicating that these duplications are detected by intragenomic comparisons only.

Consequently, the most-straightforward method to detect such ghost and/or hidden duplications is to first identify all pairs of non-hidden duplicated segments by using a tool such as ADHoRe described above. Next, it is checked for every set of detected duplicated segment pairs A-B and B-C if A-C is also detected as a non-hidden duplication. If not, then A-C can, through transitivity, be considered a 'hidden' or 'ghost' duplication. The result is a set of mutually homologous segments, referred to as a multiplicon.[43] Fig. 2 shows an example of a hidden duplication. Obviously, there can be more than three segments in a multiplicon. In fact, the number of segments in a multiplicon, denoted as the multiplication level, can be important to infer the number of duplication events that have occurred in a genome's past. Indeed, if in a given genome a chromosomal segment appears in *n*-fold then a lower bound for the number of duplications that have occurred is given by $d_{\min} = \lceil \log_2(n) \rceil$ (take $\log_2$ of n and round up to the next

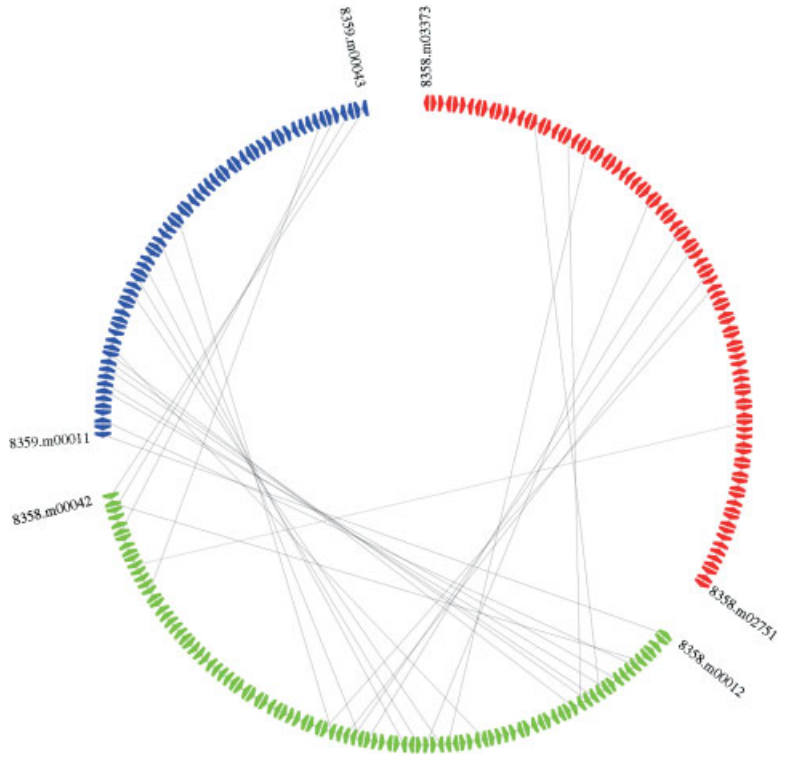

Figure 2. Example of a 'hidden' duplication in rice (*Oryza sativa*). The three segments shown are (clockwise, starting from the upper right segment): chromosome 11, loci 8358.m03373 through 8358.m02751; chromosome 11, loci 8358.m00012 through 8358.m00042 and chromosome 12, loci 8359.m00011 through 8359.m00043. Each box on a segment represents a gene. Lines between segments indicate pairs of homologous genes. Note that, while the first segment has numerous homologs with the second and the second with the third, not a single homologous gene pair is found between the first and third segment.

integer), whereas the upper bound is given by $d_{max} = n - 1$. Based on the parsimony principle, and assuming that all segments of the multiplicon have been detected, this lower bound number probably reflects the true number of segmental duplications that have occurred.

By inferring all multiplicons for the *Arabidopsis* genome, we found many multiplicons with multiplication levels (the number of segments in a multiplicon) of 5 up to 8 distributed over all five chromosomes.[43] This pattern suggests that *Arabidopsis* is likely to have undergone three rounds of duplication but not more. Additional support for three rounds of genome duplications was provided by a dating analysis where calculating the age based on the rate of synonymous substitutions (K_s) for all duplicated gene pairs in all detected duplicated segments yielded three distinct age classes.

Coping with differential gene loss

Another strategy for uncovering duplicated segments that have become unrecognizable due to differential gene loss is to compare gene order (or content) information from other related species. In a re-analysis of the *Saccharomyces cerevisiae* genome, Wong and coauthors[44] started from a classical GHM onto which a proximity plot was superimposed. A proximity plot differs from a GHM in that, in a GHM, a dot at position *x,y* represents the fact that genes *x* and *y* are homologs whereas in a proximity plot this signifies the fact that *x* and *y* are neighboring genes in another genome. The rationale behind this approach is that, for a pair of segments that has undergone considerable differential gene loss, pairs of genes that were neighbors in the ancestral sequence will also show up as diagonal patterns on the proximity plot (see Fig. 3). This diagonal pattern can be then be enhanced by superimposing a classical GHM. By combining a proximity plot using plasmid end data from 13 other hemiascomycete yeasts with a GHM, Wong and coworkers found that 82% of the *Saccharomyces* genome is duplicated, which is a dramatic increase in sensitivity when compared to the previously reported 50%.[22] Since again almost no overlapping duplications were found, the hypothesis of successive independent duplications could be ruled out. Using plasmid end sequences to obtain neighboring gene pairs when constructing the proximity plots, provides an elegant way of complementing a complete genome sequence with unassembled data from other genomes. Despite its elegance, this method is however limited to related species because it relies on the assumption that gene order is largely conserved between the genomes in the

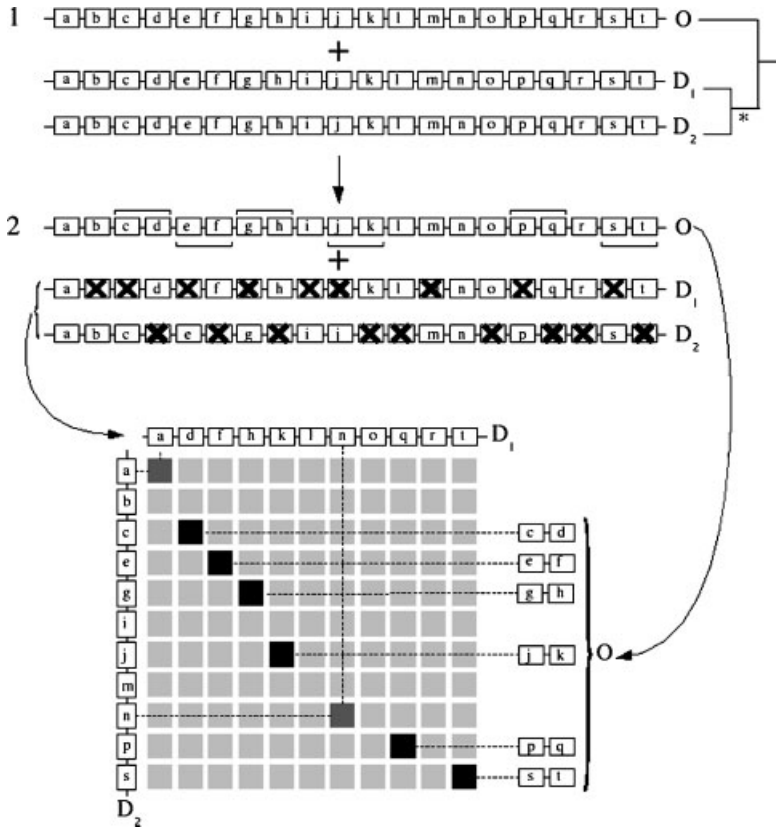

Figure 3. Principle of a proximity plot. 1. Initial situation: the genomes of two related organisms O and D are shown, each containing the same set of genes a-t. Both genomes are initially identical, except that the genome of D is duplicated (indicated by the asterisk), resulting in the chromosomes D_1 and D_2. 2. After some time, D_1 and D_2 lose a different set of genes such that the similarity between them becomes weak (only genes a and n are still present in both segments). For simplicity, it is assumed that genome O remains unchanged. When the genome sequence of genome O is only partially known (indicated by the braces above and under the segment), the proximity information (i.e. the knowledge certain genes are adjacent to each other) of genome O can be used to reveal homology between segments D_1 and D_2. This is done by constructing a dot matrix where a dot is plotted for every pair of genes from D_1 and D_2 of which the homologs in O are adjacent to each other (black cells in the matrix). Again, homology between two segments becomes apparent as a diagonal series of dots. Optionally, this proximity information can be complemented with the 'traditional' GHM-plot (dark-grey cells).

dataset. In other words, the method is not applicable to genomes that have undergone intensive rearrangements since their divergence.

Since differential gene loss is likely to occur after every large-scale duplication event, successive rounds of duplication will render it difficult to detect pairs of duplicated segments resulting from the older duplication events. Building multiplicons by detecting ghosts and/or hidden duplications as described above only partially resolves this problem. Indeed, such an approach still requires that each of the homologous segments shows significant colinearity with at least one other segment such that they can be identified by direct segment-to-segment comparison. To address this problem, different authors have proposed methods that combine information from previously identified duplicated segments in order to detect remnants from such preceding duplication events.

For instance, Blanc and coworkers[45] tried to reconstruct the ancestral *Arabidopsis thaliana* genome before the most-recent large-scale duplication event took place. To do so, they first identified all pairs of duplicated segments, using the method developed by Hokamp[37] that was already described above. Next, for every pair of duplicated segments, the age was calculated by averaging the K_s-values for all pairs of duplicated genes. The entire set of duplicated segment pairs could then be subdivided into two distinct age groups, a young one and an older one. For all younger duplicated segment pairs, the approximate ancestral gene order was then re-created by merging all genes of a segment pair into a single segment. Duplicated genes were included only once while non-duplicated genes of both segments were filled in alternately. Using this technique, an ancestral pseudogenome was created that resembles the *Arabidopsis* genome prior to the youngest duplication event. This ancestral pseudogenome was then again used to detect other duplicated regions. Within this newly obtained set of duplicated segment pairs, still a considerable amount of overlapping duplications was found, indicating that the youngest duplication event was preceded by more than one additional duplication events. The degree of overlap, however, was substantially lower than what could be expected when each duplicated segment pair was the result of an independent, local segmental duplication event. Therefore the authors concluded that the overlapping duplicated segments detected in the reconstructed pseudogenome are, at least partially, likely the result of more ancient polyploidy-type events. These results are congruent with the three large-scale duplicated events proposed by our group.[43] A similar strategy was used by Bowers and coworkers[46] who also observed evidence for three whole-genome duplication events in *Arabidopsis*.

Building profiles

The methods of Blanc and coworkers[45] and Bowers and coworkers[46] merge the gene order of two segments to detect highly diverged duplicated segments. However, it is still possible that the combined information from two segments is not sufficient to uncover highly degenerated homologous segments. This limitation is overcome in a new software tool called i-ADHoRe, developed recently.[47] This algorithm uncovers chromosomal segments that are homologous to others but can no longer be identified as such due to extreme gene loss. This is done by aligning clearly colinear segments and using this alignment as a 'genomic profile' that combines gene content and order information from multiple segments to detect these heavily degenerated homology relationships.

The approach works as follows. First, all level 2 multiplicons are identified with the previously described ADHoRe algorithm. Next, for each set of homologous segments, an alignment is created where the anchor points of the multiplicon are positioned in the same columns. Using this alignment as a profile, a new type of homology matrix can be constructed in which the rows correspond again to the positions of the gene products in their respective gene lists but where the columns correspond this time to specific positions of the profile. Once this homology matrix is constructed, it is again presented to the basic ADHoRe algorithm, which will again detect clusters of anchor points applying the same statistical validation method as described before. This time, however, new significant clusters will not reveal homology between two individual segments but between the two segments inside the profile (i.e. the initial level 2 multiplicon) and a third segment. Because this type of GHM combines gene content and order information of the different segments in the profile, it is possible to detect homology relationships with a third segment that could not be recognized by directly comparing any of the segments of the multiplicon individually with this third segment. If such a third segment is detected, it is added to the multiplicon, thereby increasing its multiplication level, and the corresponding profile is updated by aligning the new segment to it. The entire detection process can now be repeated with the newly obtained profile. The principle of this profile building approach is illustrated in Fig. 4.

Which method to choose?

In general, the methods discussed here can be divided into three different groups. The first group detects conservation of gene content and order between two homologous segments, while the second detects only conservation of gene content. The third category combines information from more than two different segments. The choice of method depends on the evolutionary distance and history of the genomes under study as well as the questions one wants to address. For instance, when studying a set of relatively closely related genomes, it can be expected that the intergenomic colinearity will still be well preserved. Thus, when the goal is only to identify colinear regions between such genomes, normal pairwise methods

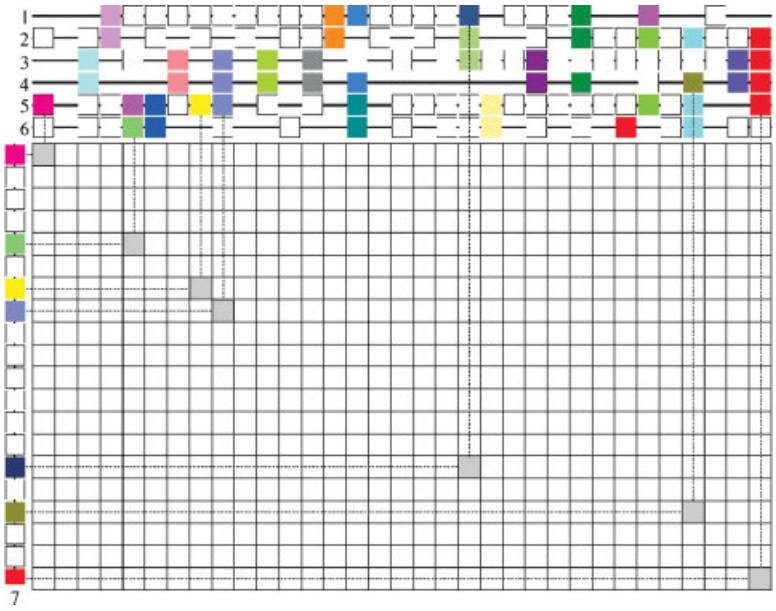

Figure 4. Detection of homology using a profile in Arabidopsis. The seven segments shown are: 1: chromosome 2, loci At2g39900 through At2g40130; 2: chromosome 3, loci At3g55750 through At3g55970; 3: chromosome 5, loci At5g05600 through At5g05820; 4: chromosome 3, loci At3g11180 through At3g11320; 5: chromosome 5, loci At5g58660 through At5g58980; 6: chromosome 3, loci At3g46950 through At3g47260; 7: chromosome 2, loci At2g38010 through At2g38240. The boxes on the segments represent genes. Genes of the same color are homologs. Tandemly duplicated genes have been remapped (not shown). Segments 1–6 (above the matrix) are aligned to form a profile. Empty positions on the segments denote gaps in the aligment. Using this profile, the homology relationship between segment 7 (left of the matrix) and the other segments in the profile/multiplicon can be established since segment 7 is clearly homologous to the profile.

such as ADHoRe,[27] FISH[28] or LineUp[29] will suffice. Also, if one wants to investigate whether a genome has undergone a large-scale duplication event in its recent evolutionary past, these methods will probably already give a quite detailed overview of the segmental duplication landscape of a genome.[23,43,48–50]

However, when the genomes under study are distantly related or one wants to analyze older segmental duplications, other strategies might be more appropriate. The choice of method then depends on the relative prevalence of local rearrangements on one hand and the extent of gene loss after divergence on the other. Indeed, the accumulation of local rearrangements that reshuffle genes over short distances, such that only the original gene content of a genomic segment is conserved, will obliterate any existing colinearity between two homologous segments. As a consequence, the homology can then only be detected by using methods that look for conserved gene content such as the algorithm of Hokamp,[37] as was illustrated in the duplication analysis of the human genome by McLysaght and coworkers.[36] Next to rearrangements, extensive gene loss may also obscure the original colinearity signal. As discussed above, methods that combine gene order and content information from multiple segments can then be used to recover homologous segments that have undergone extensive gene loss but where the relative order of the remaining genes is more or less preserved. Including segmental homology from other genomes may additionally increase the sensitivity of the latter approaches, as was illustrated in the analyses of the yeast,[51,52] *Arabidopsis*, and rice genome.[47]

Obviously, the extent of local rearrangments or gene loss is not always known beforehand. In this case, it makes sense to use a combination of different methods in order to obtain an overview of the occurrence of both phenomena. Ultimately, it is of course possible that a combination of (local) rearrangements and gene loss has reshaped homologous segments to such an extent that their homology is no longer recognizable, even with the most-advanced implementations of the map-based approach. However, it is also possible that the sequence divergence between two homologous segments is too low to reveal any differences by comparing gene content and order. In that case, direct comparison of the DNA sequences using the above mentioned methods,[12–18] will be a more suitable approach. For instance, since there exists a perfect

conservation of gene order between large parts of human and mouse, Kent and coworkers used BLASTZ[18] to assess at a detailed level the extent of recent duplications, deletions and rearrangements in both genomes.[53]

Construction of the dataset

There are a few caveats when preparing a dataset to perform a map-based analysis. First, one needs to establish objective criteria to determine what level of sequence similarity between two genes is needed to consider these genes as homologs. This can be done by simply defining a BLAST e-value cutoff. However, since e-value is an asymmetric measure (i.e. the e-value of A versus B is different then B versus A), it is then best to use reciprocal hits, meaning that the e-values of A versus B and of B versus A should both be below the cutoff value. Alternatively, one can also apply more objective homology measures such as the method of Rost.[54]

Next, when constructing the GHM, the inclusion of promiscuous sequence elements should be avoided. These are mobile elements such as (retro-)transposons that can propagate themselves throughout a genome. As a consequence, such elements result in 'clouds' of dots in a GHM that can obscure a segmental homology signal. Even worse, when using a homology detection method based on gene content only, these clouds may give rise to falsely predicted homologous segments. Because such falsely predicted duplicated segments were incorporated in their analysis, Hughes and coworkers[55] concluded erroneously that transposable elements are significantly more associated with duplicated regions than with non-duplicated regions in the *Arabidopsis* genome. This analysis could be conducted correctly by first detecting all duplicated segments in a dataset from which transposable elements have been removed and afterwards investigating the presence of these elements in the delineated duplications.

Another, albeit somewhat related, issue that should be considered prior to homology detection is gene family size. Because a gene family with n members in a genome results in $n(n-1)/2$ dots on a GHM, a family of say 100 members adds already a few thousand dots to the GHM. Consequently, the presence of such large families can result in a GHM that becomes too dense (i.e. contains too many dots) such that any homology signals can again be obscured. This is especially important when using gene-content-only-based methods since the presence of such large families increases the probability of observing a set of homologous gene pairs in two given regions by chance. For a detailed discussion of this problem, we refer to Durand and Sankoff.[40] Gene family size becomes less an issue when using methods that consider both conserved gene order and gene content, since a restriction is posed on the order in which pairs of homologs are found (see above and again Durand and Sankoff[40]).

To cope with large gene families and the possible noise that they introduce, one could put a threshold on the maximum gene family size to include in the dataset. Of course, this requires that all the genes in the dataset are first grouped into gene families, which is also not self-evident.[56–58] Alternatively, one could consider only BLAST-hits below a given e-value, as mentioned above, or take for every query only the first n hits into account. Additionally, some methods also discard tandemly duplicated genes, which also reduce the effective gene family size to some extent.

Conclusion

The availability of complete genome sequences allows biologists to investigate dynamics such as rearrangements and duplications that have shaped and continue to shape genomes. An essential part of studying genome evolution is the identification of homologous segments within and between genomes. Especially when the observed similarity between such segments has been substantially degraded during the course of evolution, their identification is far from straightforward. The development of recent computational techniques derived from the so-called map-based approach has significantly facilitated the detection of highly diverged sets of homologous segments. Different methods have been developed that are either based on the detection of both conserved gene content and order or on finding conserved gene content only between pairs of genomic fragments. However, due to phenomena such as differential gene loss, pairwise comparison of segments alone is not sufficient to uncover very ancient homology relationships. To cope with such degeneracy, highly sensitive methods have been developed that combine gene order and content information from multiple segments. It can be expected that, when more genome sequences become available, these computational tools will allow biologists to obtain a comprehensive overview of the dynamics of genomes of all kinds of organisms.

Until now, these methods have been mainly applied to the study of the genome evolution of various eukaryotes, although all methods described here are equally applicable to prokaryotic genomes as well. The fact that, up to now, genome homology studies were mainly focussed on unravelling the duplication past of individual genomes, such as yeast, *Arabidopsis* and human rather than detecting homology between different genomes, is probably due to the limited availability of complete genomes of closely related organisms. Nevertheless, comparative mapping studies between closely related organisms have already provided valuable insight in the genome evolution of these organisms[1,3] although they only uncover genomic homology at a rather coarse resolution. It can therefore be expected that, as more genomes of closely related organisms are becoming available, analyses of genomic homology between more species using the methods described here will further greatly enhance our insight of genome evolution.

Acknowledgments

We thank Jeroen Raes for helpful discussions. C.S. and K.V. are indebted to the Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie for a predoctoral fellowship.

References

1. Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
2. Nadeau JH, Taylor BA. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. Proc Natl Acad Sci USA 81:814–818.
3. Gale MD, Devos KM. 1998. Comparative genetics in the grasses. Proc Natl Acad Sci USA 95:1971–1974.
4. Ranz JM, Casals F, Ruiz A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. Genome Res 11:230–239.
5. Coghlan A, Wolfe KH. 2002. Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. Genome Res 12:857–867.
6. Bourque G, Pevzner PA. 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res 12:26–36.
7. Pevzner P, Tesler G. 2003. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res 13: 37–45.
8. Sankoff D. 2003. Rearrangements and chromosomal evolution. Curr Opin Genet Dev 13:583–587.
9. Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333–341.
10. Seoighe C. 2003. Turning the clock back on ancient genome duplication. Curr Opin Genet Dev 13:636–643.
11. Durand D. 2003. Vertebrate evolution: doubling and shuffling with a full deck. Trends Genet 19:2–5.
12. Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1–GC10.
13. Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729.
14. Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res 12:656–664.
15. Bray N, Dubchak I, Pachter L. 2003. AVID: A global alignment program. Genome Res 13:97–102.
16. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731.
17. Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, et al. 2003. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res 31:3518–3524.
18. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, et al. 2003. Human-mouse alignments with BLASTZ. Genome Res 13:103–107.
19. Vandepoele K., Simillion C, Van de Peer Y. 2004. The quest for genomic homology. Curr Genomics 5:299–308.
20. O'Brien SJ, Womack JE, Lyons LA, Moore KJ, Jenkins NA, et al. 1993. Anchored reference loci for comparative genome mapping in mammals. Nat Genet 3:103–112.
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.
22. Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713.
23. Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in Arabidopsis. Science 290:2114–2117.
24. Terryn N, Heijnen L, De Keyser A, Van Asseldonck M, De Clercq R, et al. 1999. Evidence for an ancient chromosomal duplication in Arabidopsis thaliana by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Lett 445:237–245.
25. Blanc G, Barakat A, Guyot R, Cooke R, Delseny M. 2000. Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell 12: 1093–1101.
26. Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815.
27. Vandepoele K, Saeys Y, Simillion C, Raes J, Van De Peer Y. 2002. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. Genome Res 12: 1792–1801.
28. Calabrese PP, Chakravarty S, Vision TJ. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. Bioinformatics 19 Suppl 1:i74–i80.
29. Hampson S, McLysaght A, Gaut B, Baldi P. 2003. LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. Genome Res 13:999–1010.
30. Gaut BS. 2001. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. Genome Res 11: 55–66.
31. Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H. 2002. Evidence of en bloc duplication in vertebrate genomes. Nat Genet 31:100–105.
32. Larhammar D, Lundin LG, Hallbook F. 2002. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. Genome Res 12:1910–1920.
33. Friedman R, Hughes AL. 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res 11:373–381.
34. Koszul R, Caburet S, Dujon B, Fischer G. 2004. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. EMBO J 23:234–243.
35. Cavalcanti AR, Ferreira R, Gu Z, Li WH. 2003. Patterns of gene duplication in Saccharomyces cerevisiae and Caenorhabditis elegans. J Mol Evol 56:28–37.
36. McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. Nat Genet 31:200–204.
37. Hokamp K. 2001. A Bioinformatics Approach to (Intra-)Genome Comparisons. (PhD Thesis).
38. Coulier F, Popovici C, Villet R, Birnbaum D. 2000. MetaHox gene clusters. J Exp Zool 288:345–351.
39. Hampson SE, Baldi PF, Gaut BS. 2004. CloseUp: Statistical Detection of Chromosomal Homology Using Density Alone—A Comparative Analysis. Technical report.
40. Durand D, Sankoff D. 2003. Tests for gene clustering. J Comput Biol 10: 453–482.
41. Ku HM, Vision T, Liu J, Tanksley SD. 2000. Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci USA 97:9121–9126.
42. Vandepoele K, Simillion C, Van de Peer Y. 2002. Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. Trends Genet 18:606–608.
43. Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of Arabidopsis thaliana. Proc Natl Acad Sci USA 99:13627–13632.
44. Wong S, Butler G, Wolfe KH. 2002. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. Proc Natl Acad Sci USA 99:9272–9277.
45. Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res 13:137–144.
46. Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438.
47. Simillion C, Vandepoele K, Saeys Y, Van de Peer Y. 2004. Building genomic profiles for uncovering segmental homology in the twilight zone. Genome Res 14:1095–1106.
48. Seoighe C, Wolfe KH. 1999. Updated map of duplicated regions in the yeast genome. Gene 238:253–261.
49. Vandepoele K, Simillion C, Van de Peer Y. 2003. Evidence that rice and other cereals are ancient aneuploids. Plant Cell 15:2192–2202.
50. Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci USA 101:1638–1643.

51. Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428:617–624.
52. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, et al. 2004. The Ashbya gossypii Genome as a Tool for Mapping the Ancient Saccharomyces cerevisiae genome. Science 304:304–307.
53. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 100:11484–11489.
54. Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng 12:85–94.
55. Hughes AL, Friedman R, Ekollu V, Rose JR. 2003. Non-random association of transposable elements with duplicated genomic blocks in Arabidopsis thaliana. Mol Phylogenet Evol 29:410–416.
56. Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. Nature 409:847–849.
57. Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584.
58. Krause A, Stoye J, Vingron M. 2000. The SYSTERS protein sequence cluster set. Nucleic Acids Res 28:270–272.