Correlation between sequence conservation and the genomic context after gene duplication

Richard A. Notebaart1, Martijn A. Huynen1,2, Bas Teusink1,3,4, Roland J. Siezen1,3,4 and Berend Snel1,2,*

1Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen, The Netherlands, 2Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Center, The Netherlands, 3NIZO food research, Ede, The Netherlands and 4Wageningen Center for Food Sciences, Wageningen, The Netherlands

Received July 1, 2005; Revised September 9, 2005; Accepted October 4, 2005

ABSTRACT

A key complication in comparative genomics for reliable gene function prediction is the existence of duplicated genes. To study the effect of gene duplication on function prediction, we analyze orthologs between pairs of genomes where in one genome the orthologous gene has duplicated after the speciation of the two genomes (i.e. inparalogs). For these duplicated genes we investigate whether the gene that is most similar on the sequence level is also the gene that has retained the ancestral gene-neighborhood. Although the majority of investigated cases show a consistent pattern between sequence similarity and gene-neighborhood conservation, a substantial fraction, 29–38%, is inconsistent. The observation of inconsistency is not the result of a chance outcome owing to a lack of divergence time between inparalogs, but rather it seems to be the result of a chance outcome caused by very similar rates of sequence evolution of both inparalogs relative to their ortholog. If one-to-one orthologous relationships are required, it is advisable to combine contextual information (i.e. gene-neighborhood in prokaryotes and co-expression in eukaryotes) with protein sequence information to predict the most probable functional equivalent ortholog in the presence of inparalogs.

INTRODUCTION

Comparative genomics has become an important research area in the wake of the large number of sequenced genomes that have become available in recent years. The comparison of genomes is of great importance in the prediction of gene function and to study the evolution of genomic properties such as gene-neighborhood. These comparative studies rely on homology and increasingly on orthology, because orthologs originated from a single gene in the last common ancestor by speciation (1).

The bidirectional best hit (BBH) method is a widely used homology based procedure for orthology that, in general, results in a single gene in one genome being predicted to be the ortholog of a single gene in the other genome. The BBH method has been applied in various function prediction studies, such as the construction of a conserved co-expression network and the prediction of regulatory motifs (2,3). However, one major complication in the BBH method exists when gene duplication events have occurred after the speciation of the two genomes under investigation. To distinguish these gene duplicates from more ancient and hence presumably more functionally diverged duplicates, Sonnhammer and Koonin (4) coined the phrase inparalogs for duplicated genes after a speciation event. Both these inparalogs are then orthologous to a single gene in the other species (referred as co-orthology). By only using BBH many of such truly orthologous relations will not be detected, and hence computational methods have been developed to include inparalogs (5,6). Despite the availability of methods that include inparalogs, scientists often use these programs in such a way that genes (still) only have one ortholog, thereby effectively resorting to BBH-like heuristics and ignoring small differences in sequence similarity of inparalogs to their ortholog (7). The intuitive idea behind this approach seems logical, because duplication can lead to a differentiation process in which only one of the two inparalogs retains the ancestral function. It is expected that in such cases the most similar inparalog (to the single-ortholog) is the one that has retained that ancestral function. However, especially for small differences in
sequence similarity in a large sequence space, this intuitive idea may not be correct: both inparalogs, or even only the less similar one, might carry out the ancestral function.

In order to study whether or not inparalogs have retained the (ancestral) function, we can use methods that map gene function on a genome-wide scale, such as co-expression or genomic context methods (8–10). Although these methods are indicators of biological process (11), rather than molecular function, it gives useful insights into the function of recently duplicated genes. For example, in a study that measured the conservation of co-expression, inparalogous genes were detected to often have diverged in terms of their co-expression: one of the duplicates retained co-expression, while the other did not (12). Here, we use gene-neighborhood conservation as the genomic context method, to study the relationship between gene function and sequence evolution of recent gene duplicates. Gene-neighborhood provides very strong signals for functional association between gene products within and between species (8,13–15).

In this analysis we addressed whether genes that are the most similar on the sequence level are also the ones that have retained the ancestral gene-neighborhood and hence are likely to function in the same biological process as their ortholog. Surprisingly, we have found in 29–38% of investigated co-ortholog relationships that the less similar gene pair retained the ancestral gene-neighborhood. Therefore, the BBH does not necessarily correspond to contextual information (biological process). Although, the majority of cases show a consistency between BBH and gene-neighborhood conservation, it is advisable to combine contextual information with protein sequence information to predict the most probable functional equivalent ortholog in the presence of inparalogs.

**METHODS**

**Co-orthology detection**

Our approach to investigate if there is a relationship between protein sequence similarity and gene-neighborhood conservation is based on co-orthology detection by the Inparanoid algorithm, using the default settings. Inparanoid constructs co-ortholog groups by applying a specific clustering algorithm to assign inparalogs to existing BBH pairs. This clustering algorithm is based on the assumption that inparalogs are more similar to each other than to any other sequence from the other genome (5).

**Data set of genomes**

We used the Genbank genomes of species at varying phylogenetic distances, including four Archaea (Archaeoglobus fulgidus DSM 4304, Halobacterium sp. NRC-1, Methanocaldoccus jannaschii DSM 2661 and Sulfolobus solfataricus P2), four Gram-positive bacteria (Lactobacillus plantarum WCFS1, Lactococcus lactis IL1403, Bacillus cereus ATCC 14579 and Bacillus subtilis subsp. subtilis str. 168) and four Proteobacteria (Escherichia coli K12, Pseudomonas aeruginos PAO1, Helicobacter pylori 399 and Caulobacter crescentus CB15). The genome information was taken from the NCBI database (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/).

**Gene-neighborhood conservation for function prediction**

We used gene-neighborhood conservation to predict functional equivalency (with regard to biological process) between orthologs. For each ortholog pair we parsed the Genbank files to obtain two clusters of genes (one for each ortholog), consisting of three genes $S^1$ and three genes $S^2$ of the query ortholog. We then counted the total number of ortholog pairs present in both clusters, resulting in a score scaled from 0 (no conservation) to 6 (complete conservation), irrespective of the precise gene order or the relative direction of transcription.

**Dataset of co-ortholog groups and comparison of gene-neighborhood conservation**

In order to measure the evolution of gene-neighborhood relative to the sequence evolution we specifically extracted the two-to-one (2:1) co-ortholog groups from the Inparanoid outputs (for the total number of many-to-many orthologous relationships see Table 1). These groups consist of a BBH and one additional inparalog, which is the second best hit (SBH) of the unique gene from the other genome (single-ortholog) (Figure 1). Subsequently, the number of neighboring orthologs for each ortholog pair within the co-ortholog group was compared. This comparison was done for all 2:1 co-ortholog groups.

**Figure 1.** Two-to-one (2:1) co-ortholog relationship between Bacillus cereus (B.c) and Bacillus subtilis (B.s). The number of neighboring orthologs for the BBH and the SBH were compared.

RESULTS

Consistency and inconsistency between sequence similarity and gene-neighborhood conservation

To measure the evolution of inparalogs, within two-to-one co-ortholog groups, we classified and counted the occurrences of various possible evolutionary outcomes in terms of gene-neighborhood conservation and sequence similarity (Figure 2). We are specifically interested in those cases where one of the two inparalogs has lost all traces of the ancestral gene-neighborhood while the other still retains it. We expect that the inparalog that has retained the ancestral gene-neighborhood is the preferred copy for the biological process that this chromosomal gene cluster performs. Such cases raise the question how the relative sequence similarity of the inparalogs (which is an important parameter in orthology detection) relates to retaining the ancestral gene-neighborhood. In order to analyze this, we refer to the inparalog with the highest similarity as the BBH and the inparalog with the lowest similarity as the SBH (of the single-ortholog). Inconsistencies are then defined as cases where the SBH has a conserved gene-neighborhood while at the same time the BBH does not and vice versa for the consistencies. Although the majority of investigated cases show a consistent pattern between sequence similarity and gene-neighborhood conservation, a substantial fraction, 29–38%, does not (Figure 3).

Although the number of observed inconsistencies varies between the several genome-comparison datasets, the percentages of inconsistencies are similar. Therefore, inconsistencies are not limited to specific genome comparisons. Furthermore, consistencies and inconsistencies are also found in which both BBH and SBH have a conserved gene-neighborhood, but with unequal numbers (Figure 2, class 3 and 4). Because the total number of such cases is lower, the variance in the percentage of inconsistencies is larger: 13–61%.

Inconsistencies are not caused by inparalog detection artifacts

It is known that relative BLAST hits do not necessarily reflect the actual evolutionary history of genes. One type of event that is likely to be a problem is an ancient gene duplication which has taken place before speciation, resulting in what is referred to as outparalogs (4). It is possible that BLAST hit driven methods call two genes co-orthologous to a single-ortholog (thus inparalogs), while in fact, according to phylogenetic tree
reconstruction, one of them is an outparalog and the other is a real one-to-one ortholog (Figure 4). This kind of difference between relative BLAST hits and trees are to a certain level expected because both methods are based on different approaches. For example, Inparanoid (BLAST hit driven approach) has been specifically designed for large-scale pairwise genome comparisons, based on relatively little sequence information. In contrast, more sequence information from multiple species is used in phylogenetic tree reconstruction, but reliable (large-scale) automatic procedures are not yet available. We performed several phylogenetic tree reconstructions to check if our cases of inconsistency could possibly be due to 'relative BLAST hit' artifacts. COG, MUSCLE and PHYML were used for the phylogenetic tree reconstructions (for details see Figure 4) (6,16,17). We tested an equal number of randomly selected consistencies and inconsistencies from the total dataset to confirm whether the recent gene duplicates are also inparalogs according to the phylogenetic tree. The number of confirmed inparalogs from the test set appeared to be high (85%) and, more importantly, almost equally distributed among consistent and inconsistent cases (9 and 8 out of 10, respectively). The inconsistencies are thus not likely to be the result of deficiencies of our applied large-scale method (Inparanoid).

**Inconsistencies are not caused by a lack of divergence time**

To analyze the possible cause of the inconsistencies, we have investigated the percentage of sequence identity between in paralogs within consistencies and inconsistencies. This sequence identity is a measure of how recent the duplication events are. If the inparalogs are highly identical on the amino acid sequence level, the inconsistencies could be observed because of a chance difference owing to a lack of sufficient divergence time. The analysis reveals that our observation of inconsistencies is not the result of this particular explanation, because the majority of the inparalogs are diverged to a high extent (Figure 5A). In fact, for the few cases of recently duplicated genes the ratio between consistency and inconsistency is closer to 50%, supporting the prediction that a lack of sufficient divergence time between the duplicates can cause inconsistencies by chance. The observed high level of divergence between duplicated genes could suggest a possible functional differentiation with respect to their biological process and/or molecular function (18).

**Sequence evolution of inparalogs relative to their single-ortholog**

An interesting question regarding the inparalogs is whether or not gene duplicates are differentiated with respect to their molecular function, but more importantly, if this is different for consistencies and inconsistencies. Although no direct measurement of molecular function is available on a genomic scale, we can indirectly investigate this from the sequence divergence of inparalogs relative to their single-ortholog in the other genome. Figure 5B and C shows that within the consistent and inconsistent cases both inparalogs are of comparable similarity to their single-ortholog. We thus observe little asymmetry in the rate of sequence evolution between gene duplicates, which has also been described in previous studies (18,19). As Inparanoid is not, in general, able to detect inparalogs with large sequence differences (to their single-ortholog), a small difference is to a certain level to be expected. However, we observed a tendency towards an even smaller divergence time. The analysis reveals that our observation of inconsistencies is not the result of a lack of sufficient divergence time. The analysis reveals that our observation of inconsistencies is not the result of this particular explanation, because the majority of the inparalogs are diverged to a high extent (Figure 5A). In fact, for the few cases of recently duplicated genes the ratio between consistency and inconsistency is closer to 50%, supporting the prediction that a lack of sufficient divergence time between the duplicates can cause inconsistencies by chance. The observed high level of divergence between duplicated genes could suggest a possible functional differentiation with respect to their biological process and/or molecular function (18).

![Figure 4](http://nar.oxfordjournals.org/) Example of an outparalog relationship. Two paralogs from *B. subtilis* (*mmgD* and *citZ*) are inparalogs relative to *prpC* of *E. coli*, according to Inparanoid. However, the phylogenetic tree shows an outparalog relationship between *mmgD* and *citZ*, because the duplication event took place before speciation to *B. subtilis* and *E. coli*. Therefore, *citZ* is not orthologous to *prpC*. To construct the phylogenetic tree, we first selected from the COG database the specific COG to which *mmgD*, *citZ* and *prpC* belong (6). All homologs, included in this specific COG, were used as input for MUSCLE to construct a multiple sequence alignment (16). Finally, the phylogenetic tree was constructed, from the multiple sequence alignment, using the PHYML algorithm (17).
examples of functional differentiation between inconsistent inparalogs that have been partly experimentally characterized.

Rfb and rff gene cluster

The E. coli genome contains the two gene clusters (operons) rfb and rff, each containing one member of the inparalogous pair rfbB and rffG (Figure 6). RfbB and RffG are known to both catalyze the same biochemical reaction (dTDP-4-dehydro-6-deoxy-D-glucose ↔ H₂O + dTDP-4-dehydro-6-deoxy-D-glucose), but are involved in two different biological processes (21). Genes in the rfb gene cluster (genes: rfbB, rfbD, rfbA and rfbC) are involved in the biosynthesis of O-specific polysaccharides, which are components of the membrane-localized lipopolysaccharide. In contrast, genes in the rff gene cluster are involved in the complex biosynthesis of enterobacteria common antigen, which is located in the outer membrane. L.lactis contains only the rfb gene cluster (genes: rmlA, cpsM, rmb and rmlC). E.coli rffG (part of the rff gene cluster) and L.lactis rmb (part of the rfb gene cluster) are BBHs, despite the fact that they are part of two different biological processes. In fact, E.coli rfbB and L.lactis rmbB (SBH pair) are the most reliable functional equivalents, because these two genes have a conserved gene-neighborhood (Figure 6, rfb gene cluster). In this example, the BBH does not represent the functional equivalent in terms of biological process.

Gab gene cluster

Both E.coli and B.cereus contain the gene gabT, which codes for a 4-aminobutyrate transaminase. This enzyme is part of the 4-aminobutyrate-degradation pathway. A neighboring gene of gabT on the E.coli and B.cereus genome is gabD, which is also necessary to degrade 4-aminobutyrate to succinate. Both genes are under a complex regulation process, induced by stress conditions, and are reported to be co-transcribed (22,23). However, gabT of B. cereus and goaG of E.coli are BBHs in a pairwise genome comparison. GoaG is not known to be involved in a biochemical reaction or pathway and it does not have conserved neighboring genes on the genome. In contrast, the gabT genes of E.coli and B.cereus are detected as a SBH pair, but they are both flanked by gabD. It is clear that the SBH, instead of the BBH, represents the most probable functional equivalents.

Prp gene cluster

The prp gene cluster in E.coli codes for enzymes involved in the methylcitrate cycle, in which propionate is degraded to pyruvate and succinate (Figure 7). The cluster contains prpC that codes for a 2-methylcitrate synthase. From biochemical studies it is known that PrpC has affinity for both acetyl-CoA and propionyl-CoA. Propionyl-CoA is converted to 2-methylcitrate by 2-methylcitrate synthase (EC 4.1.3.31), whereas acetyl-CoA is converted to citrate in the citric acid cycle by citrate synthase (EC 2.3.3.1). Moreover, it has been shown that 2-methylcitrate synthase and citrate synthase are regulated independently, because 2-methylcitrate synthase is only activated during growth on propionate, thereby supporting a functional difference (24).

However, two inparalogs of B.subtilis (citZ as part of a BBH and mmgD as SBH) are found in a comparison with E.coli. The citZ gene is annotated and experimentally verified as a citrate
One-to-one orthology versus gene-neighborhood conservation

We have studied how often the functional equivalents, according to gene-neighborhood conservation, are also the genes that are the most similar at the amino acid sequence level. Although in many cases there is conservation of gene-neighborhood between the detected BBHs (in the presence of inparalogs), it is surprisingly not true in all cases. We have found in 29–38% of the cases that only the less similar ortholog pair (SBH) has a conserved gene-neighborhood. Therefore, in a substantial fraction, the most probable functional equivalents are the genes in an SBH pair instead of the BBH.

Evolutionary tracks of gene duplicates in terms of biological process

As most of the duplication events are not recent (Figure 5A), the duplicates are expected to have differentiated to some degree as is also suggested by the differential retention of the ancestral gene-neighborhood. Specifically, it is interesting whether one of the gene duplicates is now active in a completely distinct biological process (i.e. neofunctionalization with respect to the process) or whether it is still active in the same biological process [i.e. subfunctionalization within the process (27,28)]. We therefore investigated the new genomic context of ‘inconsistent’ inparalogs in E.coli in relation to the ancestral context as defined by the single-ortholog. We took E.coli because the transcriptional organization of its genome is well characterized (e.g. operons). Seventeen
inconsistent cases were amenable to analysis. In 10 cases, both ‘inconsistent’ inparalogs are part of two different operons which are active in the same general biological process, according to the COG functional classification. In contrast, only two ‘inconsistent’ inparalogs are in a context that clearly indicates a different biological process: the members of one operon are classified in different COG functional categories compared with the members of the second operon. Most inconsistent inparalogs seem thus to have undergone subfunctionalization rather than neofunctionalization. The remaining five cases included inparalogs which were classified as a transcriptional unit consisting of one gene. We suspect that these isolated inparalogs are still functionally associated to their ancestral components, but active under specific cellular conditions. The importance of such differential regulation (activation under different conditions) of genes that catalyze similar reactions in large-scale biological networks have been shown previously by Ihmels et al. (29). We have, for example, observed individual genes, such as substrate-binding proteins of ATP-binding cassette (ABC) transporter systems, which do not reside in a gene cluster with other essential ABC transporter components. It is possible that such a solitary substrate-binding inparalog codes for a protein with slightly altered substrate specificity, but which still uses the ancestral ABC transport system to transport external substrates (e.g. Supplementary Data) under specific conditions.

Why do we observe inconsistencies?
It is very likely, and for a number of cases we have documented, that inparalogs which have lost the ancestral gene-neighborhood are not the preferred copy for their ancestral process and have subfunctionalized within the same general process or have even neofunctionalized. However, the gene duplicates with a different context can be more similar to the single-ortholog on the sequence level, which raises the questions: what does this imply for the molecular function of ‘inconsistent’ inparalogs and why do we observe these inconsistencies? As we observe little difference of inparalogs to their single-ortholog and an even smaller difference for the inconsistent inparalogs, it suggests that both inparalogs have a very similar molecular function. Nevertheless, it cannot be excluded that substantial changes in molecular function have occurred in the evolution of these inconsistent inparalogs, given that only one amino acid substitution can change, e.g. the substrate specificity of an enzyme (20). A change in the molecular function of inconsistent inparalogs could negatively contribute to the fitness of a given species (because the inparalog is still in the conserved gene-neighborhood), which would imply that more inconsistencies are observed for populations with a low or even a negative selection co-efficient. We tested this effect by comparing the inconsistency percentage to the ‘selected codon usage bias’ as a measure of the strength of selection in that species (30). To our surprise we did not find this effect. Instead there is no observable correlation between selection strength and the percentage of inconsistency, although we might have too few species to detect any such trend even if it exists (Supplementary Data). Given our inability to detect any correlation, let alone a significant negative correlation, it seems likely that most of our inparalogs have a very similar molecular function. Any inconsistencies are then a chance outcome: both duplicates have diverged, but at (roughly) the same evolutionary speed (Figure 5B and C). Such a similar rate of sequence evolution has been demonstrated previously, at least for recent gene duplicates (18), and occurs because most amino acids substitutions have only been subject to purifying selection and not to adaptive selection (20).

Implication of co-orthology on function prediction
Our large-scale analysis and the experimentally characterized cases confirm that orthology detection by BBH can negatively influence function predictions when gene-neighborhood conservation and co-orthology are not considered. The occurrence of inparalogs is an issue for at least 16% of the genes in pairwise species comparisons (Table 1), but it is even more important for group orthology schemes, such as COG, where virtually no orthologous group consists of only one gene per species. The importance of including inparalogs was also recently shown for increasing the accuracy in operon predictions (31). However, the evolutionary fate of the investigated inparalogs with respect to their ancestral function is still not completely known [several possible functional diversification scenarios do exist, as recently discussed by Hughes (27)]. Therefore, in line with the original (evolutionary) definition of orthology one should include inparalogs and take both genes as equally probable candidates for function prediction. If one insists to pinpoint a functional equivalent inparalog, gene-neighborhood should be combined with protein sequence conservation. By such a combination one should take into account that the inparalog which is located in a conserved gene-neighborhood is likely to be the preferred copy for the ancestral process. In cases where gene-neighborhood conservation is absent (like in eukaryotes) the general principle of contextual information can still be applied to increase the accuracy of function prediction. There is sufficient contextual information available for eukaryotes, e.g. co-expression datasets and predicted transcription factor-binding sites. Other types of functional information are interactions from proteomics, evolutionary conservation of gene fusion and literature mining. In fact, some of these other types of contextual information (i.e. protein–protein interactions and microarray derived co-expression) have been used recently to study the functional differentiation of duplicated genes in eukaryotes (32,33). As a result of many upcoming functional genomic studies, including genome sequencing projects as well as high-throughput co-expression and protein–protein interaction analysis, the importance of contextual information in gene function prediction will rapidly increase.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We thank Jos Boekhorst, Robert Kerkhoven, Bas Dutilh and Christof Francke for stimulating discussions and we thank Paul Sharp for providing unpublished data on the strength of selected codon usage bias for archaea. This work was part of

6170  Nucleic Acids Research, 2005, Vol. 33, No. 19
REFERENCES